Kernels Semàntics per a Clustering de Patrons*

Edgar Gonzàlez i Pellicer

Índex

*Memòria de la recerca realitzada durant la meva estada a la New York University amb el suport d'una beca BE de la Generalitat de Catalunya

Resum

Les tècniques de clustering poden ajudar a reduir la supervisió en processos d'obtenció de patrons per a Extracció d'Informació. Tanmateix, és necessari disposar d'algorismes adequats a documents, i aquests algorismes requereixen mesures adequades de similitud entre patrons. Els kernels poden oferir una solució a aquests problemes, però l'aprenentatge no supervisat requereix d'estratègies més astutes que l'aprenentatge supervisat per a incorporar major quantitat d'informació.

En aquesta memòria, fruit de la meva estada de mes d'Abril al de Juny de 2007 al projecte Proteus de la New York University, es proposen i avaluen diversos kernels sobre patrons. Inicialment s'estudien kernels amb una família de patrons restringits, i a continuació s'apliquen kernels ja usats en tasques supervisades d'Extracció d'Informació. Degut a la degradació del rendiment que experimenta el clustering a l'afegir informació irrellevant, els kernels se simplifiquen i es busquen estratègies per a incorporar-hi semàntica de forma selectiva. Finalment, s'estudia quin efecte té aplicar clustering sobre el coneixement semàntic com a pas previ al clustering de patrons.

Les diverses estratègies s'avaluen en tasques de clustering de documents i patrons usant dades reals.

Clustering techniques can help in reducing supervision in processes of pattern acquisition for Information Extraction. However, clustering algorithms suited to documents must be available, and these algorithms require suitable pattern similarity measures. Kernels can offer a solution to these problems, but non supervised learning requires smarter strategies than supervised learning to incorporate more information.

In this report, which summarizes my stay from April to June 2007 at the Proteus Project at New York University, various kernels over patters are proposed and evaluated. Initially, kernels over a certain family of restricted patterns are studied, and then kernels which have already been used in supervised Information Extraction tasks are applied. Due to the degradation of performance experimented by clustering when adding irrelevant information, kernels are simplified and strategies for the incorporation of semantics in a selective way are sought. Finally, the effect of applying clustering over the semantic knowledge, previous to pattern clustering, is studied.

The various strategies are evaluated on document and pattern clustering tasks using real data.

1 Introducció

La meva tesi s'emmarca en el Processament del Llenguatge Natural, dins l'àrea de la Intel·ligència Artificial, i té com a eix la incorporació de tècniques de clustering a metodologies d'obtenció de patrons per a Extracció d'Informació, amb l'objectiu de reduir-ne la supervisió.

L'Extracció d'Informació busca l'obtenir automàticament informació estructurada a partir de documents textuals. Així, si es té una notícia com la de la figura 1, corresponent a un partit de basquetbol, l'objectiu és identificar-hi els equips, jugadors i entrenadors que hi apareixen, així com les relacions que existeixen entre ells, obtenint una estructura com la de la figura 2.

Els primers sistemes d'Extracció d'Informació es van desenvolupar durant la dècada dels 1970 (DeJong, 1979), però fins a la dècada dels 1990 es tractava de sistemes molt complexos i que requerien molt d'esforç humà, tant en la seva construcció com en la seva adaptació a diversos dominis. Tanmateix, a principis dels 1990 es va produir una simplificació de l'arquitectura dels sistemes d'Extracció d'Informació (Hobbs, 1993), fet que va permetre l'aplicació de mètodes d'Aprenentatge Automàtic. L'objectiu era trobar tècniques per adaptar els sistemes de manera eficaç, eficient i el menys supervisada possible (Riloff, 1993).

Encara que recentment s'han desenvolupat sistemes d'obtenció de patrons per a Extracció d'Informació que requereixen una quantitat molt petita de supervisió, com ara els basats en bootstrapping (Yangarber, 2003; Surdeanu et al., 2006), aquesta encara hi és present, sigui en la forma d'un conjunt inicial de patrons o en la forma d'una classificació dels documents d'un conjunt en rellevants i no rellevants a la tasca (figura 3a). La meva idea és reduir encara més aquesta supervisió utilitzant tècniques de clustering.

El terme clustering engloba el conjunt de tècniques que tenen com a objectiu agrupar objectes similars dins un conjunt, d'acord a una certa noció de semblança. Aquestes tècniques han estat estudiades i utilitzades en moltes àrees de l'Aprenentatge Automàtic i el Reconeixement de Patrons. Dos estudis del ventall de tècniques existents per a clustering es poden trobar a Jain et al. (1999) i Xu i Wunsch (2005).

La factibilitat de l'aplicació de tècniques de clustering a l'aprenentatge de patrons depèn clarament de l'existència de mètodes de clustering adequats. I per aquesta raó, el principal focus de recerca en una primera fase de la tesi va ésser el clustering de documents.

Els experiments aquí presentats es mouen en aquesta direcció. A la secció 2 es detallen els experiments previs sobre clustering de documents usant patrons que havíem dut a terme abans de l'estada. A partir d'aquí es descriuen els experiments realitzats a la New York University. La secció 3 inclou experiments amb clustering de documents usant kernels sobre patrons. Les següents seccions exposen la recerca més específica sobre kernels entre patrons: començant per la secció 4, que conté experiments amb kernels usats en tasques supervisades d'Extracció d'Informació; seguint per la secció 5, on s'intenta millorar els kernels amb semàntica; i acabant a la secció 6, on s'intenta reduir la complexitat de la representació semàntica usant un altre nivell de clustering. Per últim, la secció 7 extreu conclusions generals de tots els experiments realitzats durant la meva estada.

Addicionalment, i per a facilitar el seguir de forma àgil el present document, a l'apèndix A s'hi pot trobar un quadre detallat amb tota la notació utilitzada.

Victòria del Sant Hipòlit a la pista del La Gleva
Els homes del Sant Hipòlit han demostrat la seva vàlua a domicili, aconseguint un resultat de 67 - 87 davant un La Gleva en què Joan Genís no ha brillat com habitualment, i no ha passat d'uns modestos 3 punts. En canvi, el base Francesc Veguer ha aconseguit 14 punts per als Hipolitencs, i ha estat una peça determinant en la victòria del conjunt de Jaume Forns.

Figura 1: Exemple de notícia

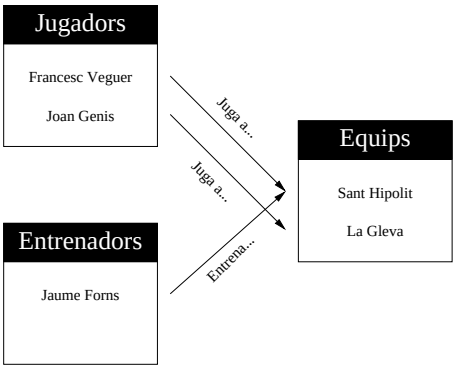


Figura 2: Informació extreta de la notícia anterior

2 Experiments Previs

Als primers experiments que havíem dut a terme sobre el procés d'aprenentatge de patrons (descrits al meu projecte de tesi (Gonzàlez, 2007), juntament amb una introducció més detallada als temes de l'Extracció d'Informació i l'obtenció automàtica de patrons) havíem utilitzat un esquema seqüencial: els documents d'una col·lecció són primerament clusteritzats, i posteriorment s'identifiquen els patrons rellevants a cada clúster (figura 3b).

Tanmateix, al projecte de tesi també havia presentat dos esquemes alternatius a l'aprenentatge de patrons usant tècniques de clustering: un de col·laboratiu en què el clustering dels documents i l'aprenentatge dels patrons intercanvien informació per a obtenir una solució final (figura 3c); i un de conjunt en què el clustering i l'aprenentatge són resultat d'un procés únic (figura 3d).

Aquest darrer esquema conjunt ja havia despertat el nostre interès abans de l'estada a la New York University (NYU). La primera aproximació que havíem estudiat usant aquest esquema conjunt havia consistit en obtenir els patrons que apareixen a cada document de la col·lecció, i utilitzar aquests patrons com a característiques per a clusteritzar els documents, en comptes de les paraules. Per a restringir l'espai de cerca dels patrons a obtenir, s'estableix un conjunt de meta-patrons a priori, de què els patrons considerats havien de ser instanciacions.

A continuació descriuré amb més detall aquesta aproximació, els experiments que vam fer amb ella i els seus resultats. L'anàlisi d'aquests resultats va ser el que va menar a la recerca que posteriorment vam dur a terme durant la meva estada.

2.1 Aproximació

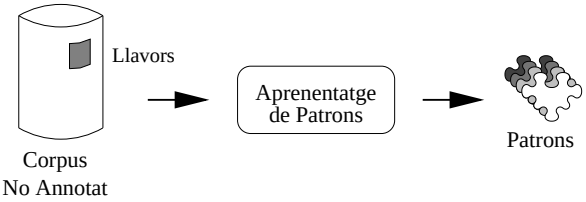
L'aproximació al clustering que presentem en aquesta secció és una adaptació a clustering del model probabilístic generatiu per a classificació de documents de Nigam et al. (2000), que ja ha estat usat per a clustering en altres treballs, com ara Surdeanu et al. (2005).

El punt de partida del mètode és una col·lecció de documents $\mathcal{D} = \{d\}$, en què cada document d es pot veure com una seqüència d'esdeveniments $d = \{\eta_1 \dots \eta_{|d|}\}$. Cada esdeveniment η_j correspon a una paraula, patró, etc. Els esdeveniments pertanyen a un vocabulari d'esdeveniments E que considerarem enumerat: $E = \{\bar{\eta}_1 \dots \bar{\eta}_{|E|}\}$.

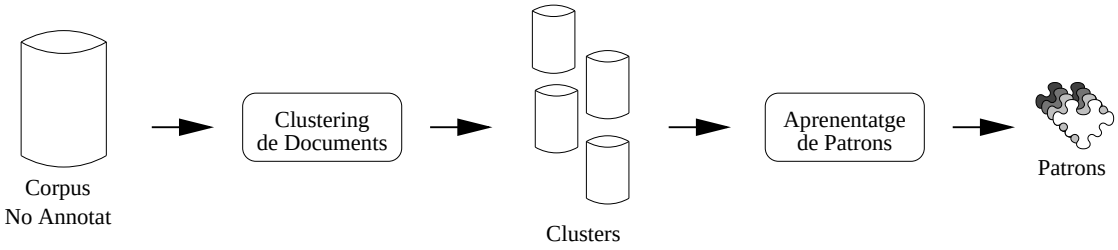
Un clustering Π és una partició del conjunt $\mathcal{D}$ en conjunts disjunts (o clústers) $\Pi = \{\pi_1 \dots \pi_k\}$. El clustering Π també es pot veure com una funció que s'aplica als documents d per a obtenir etiquetes $\{1 \dots k\}$, corresponents als clústers:

$$\begin{aligned} \Pi\colon \mathcal{D} &\rightarrow \{1 \dots k\} \\ \Pi(d) = m &\leftrightarrow d \in \pi_m \end{aligned}$$

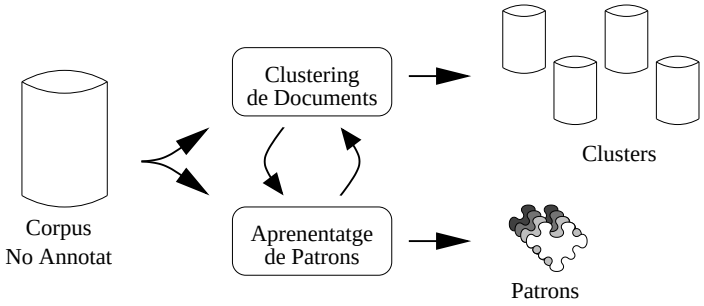
El model generatiu considerat modela la col·lecció usant una mescla (de l'anglès *mixture*) de k components, cadascun d'ells corresponent a un clúster. Dins cada component, l'aparició d'un



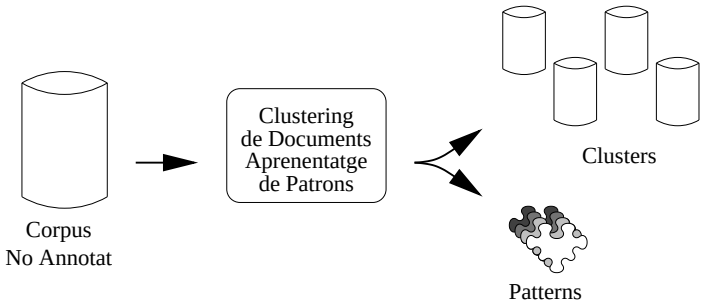
a) Aprenentatge de patrons amb selecció manual de llavors



b) Clustering i aprenentatge de patrons seqüencials



c) Clustering i aprenentatge de patrons col·laboratius



d) Clustering i aprenentatge de patrons conjunts

Figura 3: Esquemes per a l'aprenentatge de patrons d'Extracció d'Informació

esdeveniment en la seqüència es modela usant una distribució multinomial, i cada aparició es considera independent de la resta donada la component (assumpció de *Naive Bayes*). El model resultant pot formular-se com:

$$\begin{aligned} p(d \mid \Theta) &= \sum_{m=1}^{k} \alpha_m \cdot p(d \mid \Theta_m) \\ p(d \mid \Theta_m) &= \prod_{j=1}^{|d|} p(\eta_j \mid \Theta_m) \\ p(\eta_j \mid \Theta_m) &= \prod_{e=1}^{|E|} \vartheta_{me}^{\delta(\eta_j, \bar{\eta}_e)} \end{aligned}$$

on la funció d'igualtat $\delta(x, x')$ val 1 quan $x = x'$ i 0 altrament. Addicionalment, cal que els paràmetres acompleixin les restriccions:

$$\begin{aligned} \sum_{m=1}^{k} \alpha_m &= 1 \\ \forall m \sum_{e=1}^{|E|} \vartheta_{me} &= 1 \end{aligned}$$

és a dir, les $\{\alpha_m\}$ i cada conjunt de $\{\vartheta_{me}\}$ han de pertànyer a un n-simplex.

L'estimació de Màxima Varsemblança o de Màxim a Posteriori dels paràmetres $\widehat{\Theta} = \{\alpha_1, \widehat{\Theta}_1 \ldots \alpha_m, \widehat{\Theta}_M\}$ pot obtenir-se utilitzant l'algorisme d'Expectation-Maximization (Dempster et al., 1977). Donat que, com hem comentat, els paràmetres del model es troben dins un n-simplex, com a distribució a priori es pot utilitzar una distribució de Dirichlet, assignant els paràmetres del prior tots iguals a 1. En aquest cas, l'estimació de Màxim a Posteriori és equivalent a una estimació amb suavitzat de Laplace (Manning i Schütze, 1999).

Per a obtenir el clustering, es busca el component m que té la màxima probabilitat d'haver generat cada document d en aquesta estimació $\widehat{\Theta}$:

$$\begin{aligned} \Pi(d) &= \arg\max_m \left[p(m \mid d, \widehat{\Theta}) \right] \\ &= \arg\max_m \left[\alpha_m \cdot p(d \mid \widehat{\Theta}_m) \right] \end{aligned}$$

Es consideren dos tipus de característiques dels documents com a esdeveniments en el model anterior:

Mots Els mots dels documents, eliminant xifres i *stop words* i ignorant la seva caixa, com a Surdeanu et al. (2005).

Patrons Els patrons que apareixen als documents. Es tracta de patrons sintàctics instanciats a partir d'un conjunt reduït de meta-patrons (del tipus mostrat a la figura 4), i que inclouen Entitats amb Nom dels tipus de les conferències MUC (Persona, Lloc, Organització i Altres), així com Dates i Quantitats, com a Surdeanu et al. (2006). Un exemple del procés d'instanciació es pot veure a la figura 5.

En tots dos casos, les característiques que només apareixen en un document s'eliminen.

2.2 Experiments

Per tal d'estudiar l'efectivitat de les dues representacions **Mots** i **Patrons** proposades a l'hora de capturar l'estructura semàntica de col·leccions de documents, vam aplicar el model proposat a la

sv	:	Subjecte	-	Verb				
svo	:	Subjecte	-	Verb	-	Objecte		
svoc	:	Subjecte	-	Verb	-	Objecte	-	Complement
svc	:	Subjecte	-	Verb			-	Complement
so	:	Subjecte			-	Objecte		
soc	:	Subjecte			-	Objecte	-	Complement
sc	:	Subjecte					-	Complement
vo	:			Verb	-	Objecte		
voc	:			Verb	-	Objecte	-	Complement
oc	:					Objecte	-	Complement
vc	:			Verb			-	Complement

Figura 4: Meta-patrons

Francesc Veguer va anotar 20 punts a La Gleva.									
sv	(	PERSONA	,	anotar					)
svo	(	PERSONA	,	anotar	,	punts			)
svoc	(	PERSONA	,	anotar	,	punts	,	a LLOC	)
svc	(	PERSONA	,	anotar	,			a LLOC	)
so	(	PERSONA	,			punts			)
soc	(	PERSONA	,			punts	,	a LLOC	)
sc	(	PERSONA	,					a LLOC	)
vo	(			anotar	,	punts			)
voc	(			anotar	,	punts	,	a LLOC	)
oc	(					punts	,	a LLOC	)
vc	(			anotar	,			a LLOC	)

Figura 5: Exemple d'instanciació dels meta-patrons en patrons

secció anterior sobre diversos conjunts de dades, i vam avaluar la qualitat del clustering resultant comparant-lo amb una classificació dels documents en categories considerada *real*. La inicialització de l'algorisme d'Expectation-Maximization va ser manual, per a evitar que el procés d'inicialització influís en el resultat final.

2.2.1 Dades

Vam utilitzar un conjunt de 4 col·leccions de documents en llengua anglesa provinents de diverses fonts. En tots els casos es tracta de dades provinents del món real.

APW El subconjunt d'Associated Press (any 1999) de la col·lecció AQUAINT. Com a categoria vam utilitzar l'etiqueta CATEGORY de les pròpies dades.

LAT El subconjunt del Los Angeles Times de la col·lecció del TREC-5. Com a categoria vam agafar el departament del diari que va generar l'article, com a Zhao i Karypis (2004).

REU El subconjunt de 10 categories més freqüents de la col·lecció Reuters-21578. De forma similar al treball de Nigam et al. (2000) i Surdeanu et al. (2005), utilitzem la partició ModApte. Més concretament, usem la partició de test directament, ja que els algorismes a estudiar són no supervisats.

SMT Una col·lecció desenvolupada per a l'avaluació del sistema de recuperació d'informació SMART.

Col·lecció	Documents	Categories
APW	5000	11
LAT	5000	8
REU	2545	10
SMT	5467	4

Taula 1: Mida de les col·leccions

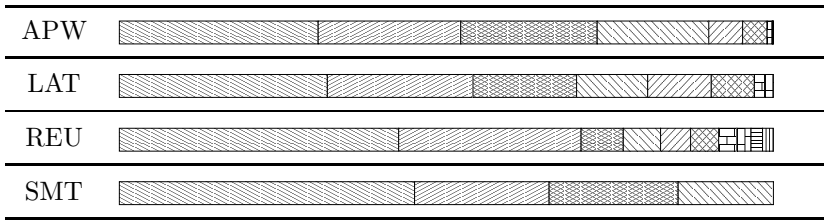


Figura 6: Distribució de les categories en les col·leccions

El nombre de documents i categories de cada col·lecció pot veure's a la taula 1, i la figura 6 conté una representació gràfica de la distribució de documents en categories dins cada col·lecció. Pot veure's com, especialment a APW, LAT i REU, existeix un conjunt de relativament poques categories que cobreix la gran majoria dels documents, mentre que altres categories tenen una presència força marginal.

2.2.2 Mètriques

Per a mesurar la qualitat del clustering resultant $\Pi = \{\pi_1 \dots \pi_k\}$, es compara el clustering amb la classificació considerada *real* dels documents en categories $\Lambda = \{\lambda_1 \dots \lambda_q\}$. Les mètriques d'avaluació utilitzades són:

Puresa (Pur) Avalua el grau en què cada clúster π_m conté documents d'un única categoria λ_c. La puresa d'un clúster és la fracció de la mida del clúster que representa la seva categoria majoritària (Zhao i Karypis, 2004). La puresa del clustering és la mitja ponderada de la puresa de cada clúster:

$$\begin{aligned} Pur(\pi_m, \Lambda) &= \frac{\max_{\lambda_c \in \Lambda} |\pi_m \cap \lambda_c|}{|\pi_m|} \\ Pur(\Pi, \Lambda) &= \frac{\sum_{m=1}^{k} |\pi_m| \cdot Pur(\pi_m, \Lambda)}{\sum_{m=1}^{k} |\pi_m|} \end{aligned}$$

Puresa Inversa (IPur) Avalua el grau en què els documents de cada categoria estan agrupats en un sol clúster. La puresa inversa d'una categoria λ_c és la fracció de la mida de la categoria que representa el clúster amb més documents d'aquesta categoria. La puresa inversa del clustering és la mitja ponderada de la puresa inversa de cada categoria:

$$\begin{aligned} IPur(\Pi, \lambda_c) &= \frac{\max_{\pi_m \in \Pi} |\pi_m \cap \lambda_c|}{|\lambda_c|} \\ IPur(\Pi, \Lambda) &= \frac{\sum_{c=1}^{q} |\lambda_c| \cdot IPur(\Pi, \lambda_c)}{\sum_{c=1}^{q} |\lambda_c|} \end{aligned}$$

F1 És la mitja harmònica de puresa i puresa inversa.

$$F_1(\Pi, \Lambda) = \frac{2 \cdot Pur(\Pi, \Lambda) \cdot IPur(\Pi, \Lambda)}{Pur(\Pi, \Lambda) + IPur(\Pi, \Lambda)}$$

	Mots			Patrons		
Col	Pur	IPur	F1	Pur	IPur	F1
APW	0.732	0.763	0.747	0.583	0.593	0.588
LAT	0.786	0.832	0.809	0.584	0.660	0.620
REU	0.804	0.833	0.818	0.690	0.705	0.697
SMT	0.934	0.934	0.934	0.613	0.617	0.615

Taula 2: Resultats del clustering de documents amb el model probabilístic

svo	(	PERSONA	,	anotar	,	punts		)
svo	(	jugador	,	anotar	,	punts		)
svo	(	PERSONA	,	anotar	,	punts		)
svo	(	PERSONA	,	aconseguir	,	punts		)
svo	(	PERSONA	,	anotar	,	punts		)
svoc	(	PERSONA	,	anotar	,	punts	, a LLOC	)

Figura 7: Exemples de parells de patrons similars

2.3 Resultats

Els resultats obtinguts es poden veure a la taula 2.

A la taula es pot observar clarament com, mentre que el clustering usant la representació **Mots** és capaç de capturar l'estructura de les col·leccions, amb valors de F1 per sobre de 0.80 en gairebé tots els casos, i fins i tot per sobre de 0.90 en la col·lecció **SMT**; el clustering usant la representació **Patrons** condueix a clusterings de qualitat inferior en totes les mètriques.

Creiem que una raó d'aquest comportament rau en el fet que el model proposat fa decisions polars respecte a la similitud d'esdeveniments: o són iguals o són diferents. En el cas de la representació **Mots** això no degrada significativament el rendiment, especialment en una llengua de morfologia simple com l'anglès. Els esdeveniment de la representació en **Patrons** porten molta més informació semàntica, però si el model no és capaç de detectar un cert grau de similitud entre parells de patrons com els de la figura 7, es produeix una gran pèrdua d'informació, i en resulta la incapacitat de detectar relacions entre documents que parlen del mateix tema.

Resulta doncs lògic de creure que millorant la funció de similitud entre patrons, aquests resultats inicials podien millorar. Amb aquesta idea vam dur a terme els diferents experiments que van tenir lloc durant l'estada de què aquest document és memòria, i que es descriuen a les seccions següents.

3 Clustering de Documents amb Kernels

Els primers experiments duts a terme durant l'estada van tenir com a objectiu el trobar una funció de similitud entre objectes en un cert espai (en el nostre cas patrons) de cara a clusteritzar documents. Aquest és un problema habitual en el camp de l'Aprenentatge Automàtic en general, i també en el del Processament del Llenguatge Natural en particular.

Una solució habitual és utilitzar kernels. Un kernel (o funció de kernel) és un producte escalar de les representacions de dos objectes en un determinat espai vectorial Φ, habitualment d'una molt alta dimensionalitat, i obtingudes usant una funció de mapeig φ (Mercer, 1909).

$$\begin{aligned} \varphi\colon X &\rightarrow \Phi \subseteq \mathbb{R}^n \\ K\colon X \times X &\rightarrow \mathbb{R} \end{aligned}$$

sv	(	PERSONA	,	anotar								)
	(	PERSONA	,	anotar	,	$\perp$	,	$\perp$	,	$\perp$		)
voc	(			anotar	,	punts	,	a LLOC				)
	(	$\perp$	,	anotar	,	punts	,	a	,	LLOC		)

Figura 8: Exemples de representació formal dels patrons

$$K(x, x') \quad = \quad \varphi(x) \cdot \varphi(x')$$

L'interès dels kernels per a l'Aprenentatge Automàtic rau en els casos en què hi ha una formulació alternativa de la funció K que no necessita calcular explícitament aquestes representacions $\varphi(x)$ i $\varphi(x')$ en l'espai vectorial d'alta dimensionalitat. Existeix un seguit de mètodes d'Aprenentatge Automàtic que depenen únicament de productes escalars entre objectes i que, per tant, poden beneficiar-se de kernels d'aquest tipus per a aprendre en espais de molt alta dimensionalitat sense incórrer en problemes d'eficiència espacial o temporal. Entre aquests mètodes podem citar:

- Perceptrons (Freund i Schapire, 1998)
- Màquines de Vectors de Suport (Vapnik, 1995)
- Anàlisi de Components Principals (Scholkopf et al., 1998)

També existeixen diverses aproximacions al clustering usant kernels. Una de les més senzilles és el kernel k-means de Girolami (2002), que és una extensió de l'algorisme de k-means clàssic (McQueen, 1967) usant kernels en el càlcul de la distància.

3.1 Aproximació

Per tal d'obtenir un kernel sobre documents per a poder-los clusteritzar, considerem en primer lloc un kernel sobre patrons. El primer pas és formalitzar els patrons usats en la secció 2. Si tenim un vocabulari de lemes L a què afegim un valor indeterminat $\perp \notin L$ per a obtenir el conjunt $L^* = L \cup \{\perp\}$, podem definir el conjunt de patrons Ξ com el conjunt de tuples ξ tals que:

$$\begin{aligned} \xi \quad &= \quad (l_s, l_v, l_o, l_p, l_c) \\ l_s, l_v, l_o, l_p, l_c \quad &\in \quad L^* \end{aligned}$$

on l_s correspon al subjecte, l_v al verb, l_o a l'objecte directe, l_p a la preposició i l_c al complement del patró. L'objectiu del valor indeterminat $\perp$ és el de permetre patrons en què alguna de les posicions no estigui especificada. A la figura 8 podem veure com es representen alguns dels patrons de la figura 5 en aquest formalisme.

Podem definir un kernel genèric K^i d'identitat entre elements de qualsevol conjunt X a què s'hi pot haver afegit un valor indeterminat $\perp$:

$$K^i(x, x') = \begin{cases} 1 & x = x' \neq \perp \\ 0 & \textit{altrament} \end{cases}$$

Usant aquest kernel d'identitat amb el conjunt de lemes L, el kernel K_Ξ entre patrons es pot definir com la suma dels kernels corresponents a cada component:

$$\begin{aligned} \xi \quad &= \quad (l_s, l_v, l_o, l_p, l_c) \\ \xi' \quad &= \quad (l'_s, l'_v, l'_o, l'_p, l'_c) \\ K_\Xi(\xi, \xi') \quad &= \quad K^i(l_s, l'_s) + K^i(l_v, l'_v) + K^i(l_o, l'_o) + K^i(l_p, l'_p) + K^i(l_c, l'_c) \end{aligned}$$

L'objectiu és quantificar la similitud parcial entre patrons que comparteixen les mateixes paraules en posicions homòlogues.

Finalment, representant els documents $d \in \mathcal{D}$ com seqüències de patrons $d = \{\xi_1 \dots \xi_{|d|}\}$, podem estendre el kernel sobre patrons K_Ξ a un kernel sobre documents $K_\mathcal{D}$:

$$K_\mathcal{D}(d', d') = \sum_{\xi \in d} \sum_{\xi' \in d'} K_\Xi(\xi, \xi')$$

Per últim, una transformació habitual és aplicar una funció polinòmica sobre un kernel per a obtenir-ne un altre. En el nostre cas, també considerem el kernel quadràtic $K^2_\mathcal{D}$:

$$K^2_\mathcal{D}(d, d') = (K_\mathcal{D}(d, d'))^2$$

3.1.1 Kernel k-means

Un cop definit un kernel entre documents, el clustering es pot obtenir utilitzant l'algorisme de kernel k-means de Girolami (2002). Aquest algorisme es basa en reformular la distància euclidiana entre dos vectors usant kernels:

$$\begin{aligned} d(x, x') &= d_\Phi(\varphi(x), \varphi(x')) \\ &= \|\varphi(x) - \varphi(x')\| \\ &= \sqrt{(\varphi(x) - \varphi(x')) \cdot (\varphi(x) - \varphi(x'))} \\ &= \sqrt{\varphi(x) \cdot \varphi(x) + \varphi(x') \cdot \varphi(x') - 2 \cdot \varphi(x) \cdot \varphi(x')} \\ &= \sqrt{K(x, x) + K(x', x') - 2 \cdot K(x, x')} \end{aligned}$$

de manera que es pot implementar un algorisme equivalent a k-means en un espai d'alta dimensionalitat sense necessitat de treballar amb les formes explícites dels vectors, usant només kernels.

Els centroides tampoc no cal que s'emmagatzemin de forma explícita, sinó que es troben implícits en els elements que pertanyen a cada clúster. Respecte al càlcul de la distància de cada element al centroide, es pot demostrar que:

$$\begin{aligned} K(x, \bar{c}_i) &= K\left(x, \frac{\sum_{x' \in c_i} x'}{|c_i|}\right) \\ &= \frac{1}{|c_i|} \sum_{x' \in c_i} K(x, x') \end{aligned}$$

Usant una estratègia de càlcul com la de Zhang i Rudnicky (2002), l'algorisme es pot implementar de forma eficient.

3.2 Experiments

Per tal d'avaluar la capacitat del kernel proposat de capturar les semblances entre els patrons en el context del clustering de documents, vam aplicar l'algorisme de kernel k-means a les col·leccions de la secció 2.2.1. Vam dur a terme experiments amb el kernel **Lineal**, $K_\mathcal{D}$, així com amb el kernel **Quadràtic**, $K^2_\mathcal{D}$. Addicionalment, com a referència vam aplicar un k-means estàndard usant distància euclidiana sobre els documents representats com a vectors de **Mots** i de **Patrons** (atòmics, com a la secció 2.1). Aquests dos darrers models són equivalents a un kernel k-means usant la identitat entre mots o patrons, respectivament, com a kernel.

En tots els casos, vam normalitzar els vectors a unitaris abans de començar el clustering. La inicialització del kernel k-means va ser de nou manual, amb les mateixes llavors que als experiments de la secció 2.2.

3.3 Resultats

Els resultats obtinguts, expressats en les mètriques de la secció 2.2.2, es poden veure a la taula 3.

	Mots			Patrons		
Col	Pur	IPur	F1	Pur	IPur	F1
APW	0.616	0.593	0.604	0.419	0.676	0.517
LAT	0.646	0.671	0.659	0.398	0.817	0.535
REU	0.721	0.625	0.670	0.612	0.769	0.681
SMT	0.820	0.820	0.820	0.377	0.744	0.500
	Lineal			Quadratic		
Col	Pur	IPur	F1	Pur	IPur	F1
APW	0.441	0.371	0.402	0.410	0.577	0.479
LAT	0.414	0.395	0.404	0.417	0.640	0.505
REU	0.597	0.554	0.575	0.582	0.645	0.612
SMT	0.403	0.540	0.457	0.369	0.733	0.491

Taula 3: Resultats del clustering de documents amb kernels

Es pot veure com els resultats usant els kernels proposats són pitjors que els obtinguts amb distàncies euclidianes, i a la vegada aquests resultats són pitjors que els obtinguts anteriorment amb el model probabilístic i que apareixien a la taula 2.

Usar la representació de **Mots** segueix essent la millor opció, tant en termes de puresa com de puresa inversa. Les representacions amb **Patrons** no capturen les categories en les dades amb la mateixa efectivitat. Dins ella, sembla que els kernels proposats no reflecteixen adequadament la similitud entre patrons, i en la major part dels casos els valors en totes les mètriques són més alts per al clustering usant la distància euclidiana (kernel identitat) que no per als kernels **Lineal** o **Quadratic**.

3.4 Anàlisi

Analitzant les possibles causes del comportament dels kernels proposats, un fet que vam constatar és que els patrons, a pesar d'haver guanyat en expressivitat respecte als experiments anteriors, segueixen ser poc flexibles: existeixen uns meta-patrons determinats que s'instancien, i les relacions que no cauen dins l'esquema d'aquests meta-patrons no es poden capturar.

Com que els meta-patrons considerats son bàsicament verbals, les relacions expressades mitjançant estructures sintàctiques que no involucrin un verb, com ara les de modificació nominal, o les que s'expressen mitjançant preposicions, en queden fora. Aquestes relacions són una fracció important del total de relacions que existeixen en els documents, i el fet de no poder capturar-les pot fer perdre també una part important de la semàntica dels documents. És per això que vam pensar en un canvi substancial en el formalisme de patrons, augmentant-ne la seva flexibilitat. L'ús de patrons i kernels utilitzats en altres treballs en tasques d'Extracció d'Informació supervisada semblava particularment adequat.

4 Clustering de Patrons amb Kernels

En el context de tasques supervisades d'Extracció d'Informació, els mètodes basats en kernels han estat aplicats amb èxit a treballs com ara Roth i Yih (2001); Zelenko et al. (2003); Zhao i Grishman (2005). Els bons resultats obtinguts per aquest darrer treball de Zhao i Grishman (2005) en la tasca supervisada d'extracció i classificació de relacions de l'avaluació ACE (ACE05), així com el fet que es tractés d'un sistema desenvolupat al grup on s'estava duent a terme l'estada van fer que consideréssim interessant d'experimentar amb els kernels que s'havien utilitzat en aquest darrer treball.

Per tal d'avaluar directament la bondat dels kernels entre patrons, sense passar per la tasca de clustering de documents, vam optar per efectuar un clustering dels patrons directament, i comparar els clústers obtinguts amb les classes de relacions de l'avaluació ACE.

4.1 Aproximació

Considerem patrons similars als del treball de Zhao i Grishman (2005).

Cada oració d'un document és una seqüència de mots $\{\omega_1 \dots \omega_n\}$, cadascun d'ells ω_i representat com una tripleta de forma f_i, tag t_i i lema l_i: $\omega_i = (f_i, t_i, l_i)$. Addicionalment, es considera informació sintàctica, representada com a arestes e dirigides d'un mot ω_{or} a un altre ω_{ds} i amb una certa funció sintàctica σ: $e = (\omega_{or}, \omega_{ds}, \sigma)$. Podem definir el conjunt d'arestes que entren a un mot ω, $in(\omega)$ i el conjunt d'arestes que en surten, $out(\omega)$:

$$\begin{aligned} in(\omega) &= \{e \mid e = (\omega_{sr}, \omega, \sigma)\} \\ out(\omega) &= \{e \mid e = (\omega, \omega_{ds}, \sigma)\} \end{aligned}$$

En les oracions existeixen entitats ν, especificades mitjançant l'abast de mots que cobreixen $\{\omega_{in} \dots \omega_{nu} \dots \omega_{fi}\}$, essent ω_{in} l'inici, ω_{nu} el nucli, i ω_{fi} el final de l'entitat; el seu tipus τ i el seu subtipus $\bar{\tau}$, dins la jerarquia definida per a l'avaluació d'Extracció d'Informació ACE: $\nu = (\omega_{in}, \omega_{nu}, \omega_{fi}, \tau, \bar{\tau})$.

Un patró ξ és llavors el context que envolta dues entitats de la mateixa oració, i s'identifica mitjançant les dues entitats, relacionades o no:

$$\begin{aligned} \xi &= (\nu_1, \nu_2) \\ \nu_1 &= (\omega_{1in}, \omega_{1nu}, \omega_{1fi}, \tau_1, \bar{\tau}_1) \\ \nu_2 &= (\omega_{2in}, \omega_{2nu}, \omega_{2fi}, \tau_2, \bar{\tau}_2) \end{aligned}$$

De cara a la definició dels kernels, definirem tres conceptes addicionals sobre un patró:

Seqüència La seqüència $seq(\xi)$ és el conjunt de mots que apareixen entre els nuclis de les dues entitats.

$$seq(\xi) = \{\omega_{1nu} \dots \omega_{2nu}\}$$

Dependència L'arbre de dependència $dep(\xi)$ és el conjunt d'arcs de dependència sintàctica que connecten els nuclis de les dues entitats.

$$\begin{aligned} dep(\xi) = \{e_1 \dots e_n \quad \mid \quad & e_i = (\omega_{isr}, \omega_{ids}, \sigma_i) \\ \wedge \quad & (\omega_{1sr} = \omega_{1nu} \vee \omega_{1ds} = \omega_{1nu}) \\ \wedge \quad & (\omega_{nsr} = \omega_{2nu} \vee \omega_{nds} = \omega_{2nu}) \\ \wedge \quad & \forall i : (\omega_{isr} = \omega_{(i+1)sr} \vee \omega_{isr} = \omega_{(i+1)ds} \\ & \vee \omega_{ids} = \omega_{(i+1)sr} \vee \omega_{ids} = \omega_{(i+1)ds}) \\ \wedge \quad & (i \neq j \rightarrow e_i \neq e_j)\} \end{aligned}$$

Enllaç L'enllaç $enll(\xi)$ és el conjunt de mots que apareixen entre els nuclis de les dues entitats i que, a més a més, formen part de l'arbre de dependència entre els nuclis de les dues entitats:

$$enll(\xi) = seq(\xi) \cap \{\omega \mid \exists e_i \in dep(\xi) : \omega = \omega_{isr} \vee \omega = \omega_{ids}\}$$

Donades aquestes definicions, definim un conjunt de kernels bàsics, basats en el kernel d'identitat K^i definit a la secció 3.1:

Mots Kernel bàsic entre mots:

$$K_\Omega(\omega, \omega') = K^i(f, f') + K^i(t, t') + K^i(l, l')$$

Bigrames Kernel bàsic entre bigrames (parelles de mots consecutius):

$$\begin{aligned} K_{\Omega\times\Omega}((\omega_i,\omega_{i+1}),(\omega'_i,\omega'_{i+1})) &= K^i((f_i,f_{i+1}),(f'_i,f'_{i+1})) + K^i((t_i,t_{i+1}),(t'_i,t'_{i+1})) \\ &+ K^i((l_i,l_{i+1}),(l'_i,l'_{i+1})) \end{aligned}$$

Arestes Kernel bàsic entre arestes:

$$K_E(e,e') = K_\Omega(e_{sr},e'_{sr}) + K_\Omega(e_{ds},e'_{ds}) + K^i(\sigma,\sigma')$$

I a partir d'aquí definim una sèrie de kernels entre dos patrons, els mateixos que Zhao i Grishman (2005), tenint en compte diferents tipus d'informació:

Argument Informació morfològica i de tipus de les dues entitats relacionades.

$$\begin{aligned} K_N(\nu,\nu') &= K_\Omega(\omega_{nu},\omega'_{nu} + K^i(\tau,\tau') + K^i(\bar{\tau},\bar{\tau}') \\ K_{arg}(\xi,\xi') &= K_N(\nu_1,\nu'_1) + K_N(\nu_2,\nu'_2) \end{aligned}$$

Enllaç Informació morfològica de l'enllaç:

$$\begin{aligned} enll(\xi) &= \{\omega_1 \ldots \omega_n\} \\ enll(\xi') &= \{\omega'_1 \ldots \omega'_{n'}\} \\ K_{enll}(\xi,\xi') &= \sum_{\min(n,n')} K_\Omega(\omega_i,\omega'_i) \end{aligned}$$

Dependència Informació sintàctica de l'arbre de dependències:

$$K_{dep}(\xi,\xi') = \sum_{e\in dep(\xi)} \sum_{e'\in dep(\xi')} K_E(e,e')$$

Local Informació sintàctica local a les entitats relacionades (arcs de dependència entrants i sortints als nuclis de les entitats):

$$\begin{aligned} K_{loc}(\xi,\xi') &= \textstyle\sum_{e\in in(\omega_{1nu})} \sum_{e'\in in(\omega'_{1nu})} K_E(e,e') \\ &+ \textstyle\sum_{e\in out(\omega_{1nu})} \sum_{e'\in out(\omega'_{1nu})} K_E(e,e') \\ &+ \textstyle\sum_{e\in in(\omega_{2nu})} \sum_{e'\in in(\omega'_{2nu})} K_E(e,e') \\ &+ \textstyle\sum_{e\in out(\omega_{2nu})} \sum_{e'\in out(\omega'_{2nu})} K_E(e,e') \end{aligned}$$

Bigrama Informació morfològica dels bigrames que apareixen a la seqüència:

$$\begin{aligned} seq(\xi) &= \{\omega_1 \ldots \omega_n\} \\ seq(\xi') &= \{\omega'_1 \ldots \omega'_n\} \\ K_{big}(\xi,\xi') &= \sum_{i=1}^{n-1} \sum_{i'=1}^{n'-1} K_{\Omega\times\Omega}((\omega_i,\omega_{i+1}),(\omega'_{i'},\omega'_{i'1+1})) \end{aligned}$$

La principal diferència de la nostra aproximació respecte a Zhao i Grishman (2005) és que en aquest darrer treball s'usen arcs GLARF (Meyers et al., 2001) per a expressar la sintaxi, mentre que els nostres patrons usen dependències sintàctiques estàndard.

De cara al clustering dels patrons, utilitzem sis combinacions d'aquests kernels, que es poden veure a la taula 4, i que són les mateixes emprades per Zhao i Grishman (2005), raó per què les designem com a K^Z_*

K_A^Z	K_{arg}
K_B^Z	$K_A^Z + K_{enll}$
K_C^Z	$K_B^Z \cdot K_B^Z$
K_D^Z	$K_C^Z + K_{dep} + K_{loc}$
K_E^Z	$K_D^Z + K_{big}$
K_F^Z	$K_A^Z + K_{dep} + K_{loc}$

Taula 4: Kernels usats

4.2 Experiments

Com hem comentat prèviament, per a estudiar l'efectivitat dels kernels proposats, vam dur a terme una sèrie d'experiments de clustering de patrons.

Per a obtenir el màxim teòric assolible, així com per a eliminar el problema de la inicialització que tenen els algorismes de clustering iteratiu com ara el k-means, vam optar per partir de conjunts de dades correctament clusteritzats (és a dir, amb cada element assignat a la seva categoria) i aplicar-hi l'algorisme de kernel k-means amb els diversos kernels.

4.2.1 Dades

Com a conjunt de dades per als experiments vam usar el corpus de l'avaluació ACE 2005 per a la llengua anglesa. En concret, vam utilitzar les dades anotades per a la tasca de RMD (Relation Mention Detection, Detecció de Mencions de Relacions).

Com a objectes a clusteritzar considerem els parells d'entitats relacionades (no tenim en compte parells d'entitats no relacionades). El nombre resultant de parelles és de 8625.

Com a categories considerem els 18 subtipus d'entitats ACE, disposats en una jerarquia de 6 tipus, i que es poden veure a la taula 5. De forma similar a Zhao i Grishman (2005), distingim dins cada subtipus els casos en què la relació va de l'entitat que apareix abans al text a la que apareix després (d'esquerra a dreta, >) i aquells en què va en sentit contrari (de dreta a esquerra, <). L'excepció són les relacions de les relacions de tipus social, que es consideren simètriques[1]. El nombre final de categories en què clusteritzar és de 33.

4.2.2 Mètriques

Les mètriques usades per a l'avaluació de la qualitat del clustering són les de puresa, puresa inversa i F1 ja presentades a la secció 2.2.2.

4.3 Resultats

La taula 6 mostra els resultats obtinguts en la tasca de clustering usant els kernels presentats a la secció 4.1. Addicionalment, com a referència s'hi inclouen els resultats obtinguts per Zhao i Grishman (2005) en la tasca d'Extracció de Relacions sobre el mateix corpus ACE, usant les mètriques de Precisió, Recall i F1, habituals per a la tasca. Els millors valors en cada mètrica estan destacats en negreta.

Com es pot veure, així com en la tasca supervisada d'Extracció de Relacions la incorporació de més informació al kernel és beneficiosa i ajuda a millorar els resultats, per a la tasca no supervisada de clustering el rendiment es degrada a mida que augmenta la complexitat del kernel, i els millors resultats s'obtenen amb el kernel K_A^Z, que només té en compte informació morfològica i de tipus de les dues entitats relacionades.

[1] A l'avaluació ACE, les relacions de Physical també es consideren simètriques, però per als nostres experiments vam optar per mantenir-ne la asimetria

Tipus	Subtipus	Categories	
Artifact	User-Owner-Inventor-Manufacter	>Use	<Use
Gen-Affiliation	Citizen-Resident-Religion-Ethnicity	>Cit	<Cit
	Org-Location	>Org	<Org
Org-Affiliation	Employment	>Emp	<Emp
	Founder	>Fou	<Fou
	Ownership	>Own	<Own
	Student-Alum	>Stu	<Stu
	Sports-Affiliation	>Spo	<Spo
	Investor-Shareholder	>Inv	<Inv
	Membership	>Mem	<Mem
Part-Whole	Artifact	>Art	<Art
	Geographical	>Geo	<Geo
	Subsidiary	>Sub	<Sub
Person-Social	Business	Bus	
	Family	Fam	
	Lasting-Personal	Las	
Physical	Located	>Loc	<Loc
	Near	>Nea	<Nea

Taula 5: Tipus de relacions a l'avaluació ACE

4.4 Anàlisi

Quina és la causa d'aquest comportament? Nosaltres creiem que la clau està el diferent comportament que tenen els mètodes basats en kernels davant les característiques irrellevants en contextos d'aprenentatge supervisat i no supervisat. En un context supervisat, com ara el de Zhao i Grishman (2005), el fet de conèixer les categories i quins objectes pertanyen a cadascuna permet identificar aquestes característiques irrellevants (aquelles que no ajuden a distingir els objectes d'una categoria de la resta) i el procés d'aprenentatge pot ser-hi robust. En canvi, en un context no supervisat, la incorporació d'informació irrellevant canvia les distàncies i les relacions entre els objectes, portant a un clustering que pot tenir poca o cap mena de relació amb l'original.

Una forma d'analitzar els resultats de la secció 4.3 és utilitzar el que hem anomenat *Diagrama de Confusió*. Es tracta d'una representació gràfica de les diverses categories existents al corpus, i de com els objectes d'una categoria van a parar a una altra, confonent-se amb elles (d'aquí el nom). Un exemple, corresponent als resultats obtinguts a la secció anterior amb l'aplicació del k-means amb kernel **Argument** sobre les dades RMD de l'ACE, es pot veure a la figura 9.

En el diagrama es poden observar les diferents categories com a nodes, unides per arestes en forma de fletxa si hi ha documents de la categoria d'origen que han acabat clusteritzats amb els de la categoria destí. El gruix de la fletxa indica la fracció d'elements de la categoria origen que han anat a parar a la de destí, mentre que el color dels nodes només serveix per a agrupar els subtipus de relació del mateix tipus en la jerarquia ACE.

D'entre totes les confusions que apareixen al diagrama, val la pena destacar uns quants exemples que apareixen dins els tipus de relacions Social i Org-Affiliation, i que apareixen a la taula 7.

Com es pot veure, en tots aquests casos existeix una confusió entre diversos subtipus de relació dins el mateix tipus. I en tots els casos també, hi ha una de les paraules de la relació que en determina el tipus, que és la que es troba subratllada a la Taula. Una possible explicació a aquesta confusió és la manca de semàntica en el procés esbossat fins ara: la única valoració que es fa dels mots en els kernels proposats és si la seva forma, lema i categoria gramatical són les mateixes o no. Això fa que, per exemple, els diversos tipus de relacions familiars, marcades per

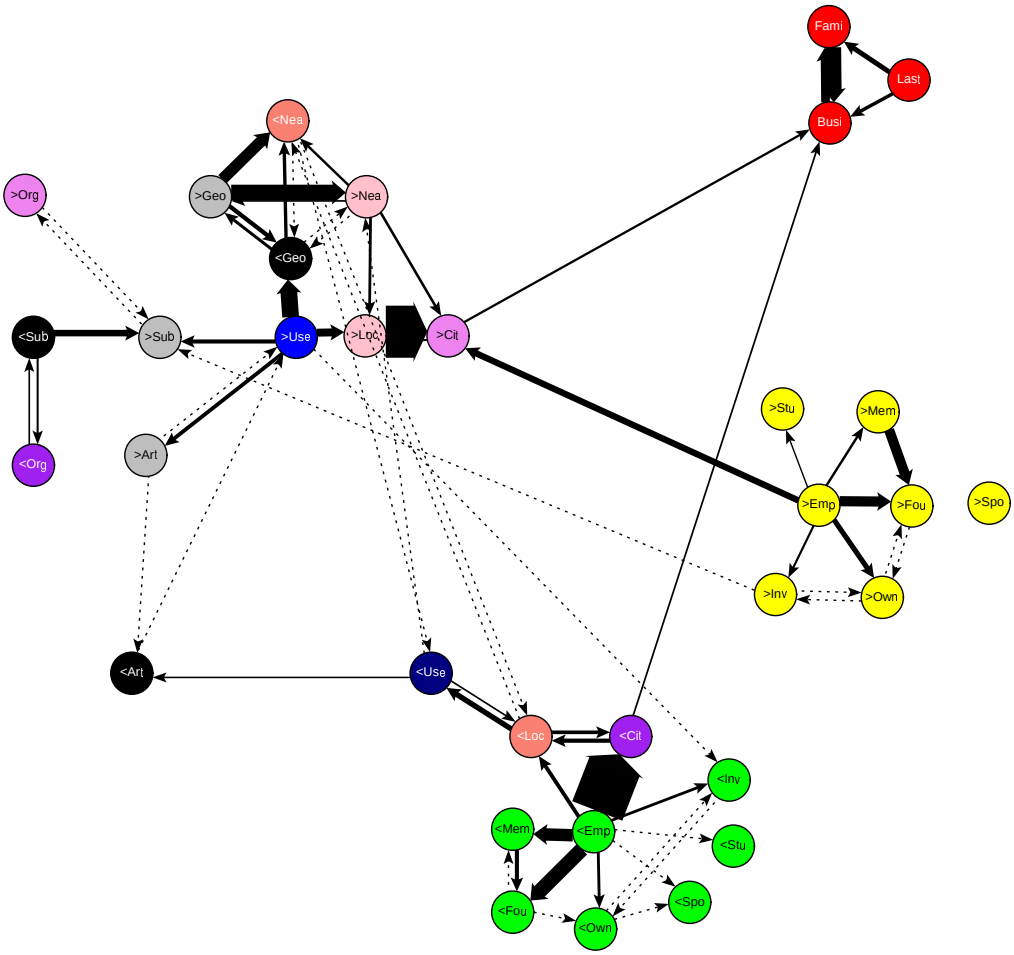


Figura 9: Diagrama de confusió per al clustering de RMD-ACE

	Clustering			Extracció		
Kernel	Pur	IPur	F1	Prec	Rec	F1
K_A^Z	**0.622**	**0.460**	**0.529**	.530	.585	.556
K_B^Z	0.482	0.251	0.330	.588	.713	.644
K_C^Z	0.470	0.315	0.377	.670	.703	.686
K_D^Z	0.432	0.275	0.336	.691	**.714**	.702
K_E^Z	0.398	0.242	0.301	**.692**	.705	**.704**
K_F^Z	0.420	0.215	0.284	.579	.685	.627

Taula 6: Resultats del clustering de patrons sobre el corpus ACE

Social			
... You are **my** best **friend** ...	`Las`	→	`Fam`
... Deirdre as **your relative** ...	`Fam`	→	`Las`
... I visited all **their families** ...	`Fam`	→	`Bus`
Org-Affiliation			
... The **head** of **AIG** ...	`>Emp`	→	`>Own`
... a **professor** of psychiatry at **NYU** ...	`>Emp`	→	`>Stu`
... the **president** of a **group** of Iraqi exiles ...	`>Mem`	→	`>Fou`

Taula 7: Exemples de confusió

paraules clau com *parents*, *father*, *mother*... no es puguin relacionar directament. En aquest casos, sembla raonable pensar que es fa imprescindible disposar d'algun tipus d'informació semàntica.

A la vista d'aquestes dades, vam veure que les estratègies utilitzades per a Extracció d'Informació supervisada no havien de ser necessàriament vàlides per a tasques no supervisades, i ens vam disposar a, partint del kernel K_A^Z, senzill i que ofereix ja al voltant del 50% de F1, incorporar informació de forma selectiva per a millorar aquests resultats i obtenir funcions de similitud entre patrons més efectives. Intentant solucionar els casos com els presentats a la taula 7, els nostres següents experiments van tenir com a objectiu la incorporació d'informació semàntica a la informació bàsica sobre les entitats relacionades.

5 Incorporació de Semàntica als Kernels

Existeixen força treballs respecte a l'aplicació de semàntica a tasques de Processament del Llenguatge Natural. D'entre els recursos que s'utilitzen en aquest camp, un dels estàndards *de facto* és la base de dades lèxica WordNet (Miller et al., 1990; Fellbaum, 1998). En ella, els conceptes s'identifiquen a partir de conjunts de mots sinònims (synsets). WordNet conté informació sobre a quins synsets pertany cada mot (sentits dels mot), així com un gran nombre de relacions entre synsets, entre què es troben les d'hiperonímia i les de la seva inversa, l'hiponímia. A partir de les relacions d'hiperonímia sobre els synsets de noms i verbs es poden definir sengles taxonomies jeràrquiques.

Existeixen diversos treballs que utilitzen WordNet i, en particular, les relacions d'hiperonímia, en el camp de l'Extracció d'Informació, com ara Turmo i Rodríguez (2002), o Culotta i Sorensen (2004), aquest darrer a més a més usant mètodes de kernels. Per aquesta raó, vam decidir seguir aquesta línia per a incorporar semàntica als kernels considerats als nostres experiments.

5.1 Aproximació

Per tal d'incorporació d'informació semàntica als kernels, es consideren nous kernels sobre mots que, a part de comparar la forma, lema i categoria gramatical de cada mot, tenen en compte els seus tancaments respecte la relació d'hiperonímia.

Formalment, si tenim el conjunt de mots $L = \{l\}$, identificats pels seus lemes, i el conjunt de synsets $S = \{s\}$, a WordNet es poden trobar els sentits d'un determinat lema l, representats mitjançant synsets:

$$sentits(l) \subset S$$

La relació d'hiperonímia existent a WordNet és defineix entre parells de synsets i és directa: $hyperd(s_1, s_2)$ si i només si el synset s_1 és un hiperònim directe del synset s_2. Aquesta relació d'hiperonímia directa $hyperd(s_1, s_2)$ es pot tancar de forma transitiva per a obtenir una relació general d'hiperonímia $hyper(s_1, s_2)$:

$$hyper(s_1, s_2) \longleftrightarrow s_1 = s_2 \vee (\exists s : hyperd(s_1, s) \wedge hyper(s, s_2))$$

La relació general d'hiperonímia és una relació d'ordre parcial definida sobre el conjunt de synsets S i, per tant, hi defineix una jerarquia. Podem definir nivells en aquesta jerarquia: els synsets que no tenen cap hiperònim directe pertanyen al nivell 1; els synsets que tenen un hiperònim directe de nivell 1 són de nivell 2; i així successivament. Definim els conjunts Top_n com els synsets que pertanyen al nivell n o menor:

$$\begin{aligned} Top_1 &= \{s \mid \nexists s' : hyperd(s', s)\} \\ Top_n &= Top_{n-1} \cup \{s \mid \exists s' \in Top_{n-1} : hyperd(s', s)\} \end{aligned}$$

Abusant de la notació, denotarem el conjunt d'hiperònims d'un synset s com $hyper(s)$, i el conjunt d'hiperònims de nivell n o menor d'un synset s com $hyper_n(s)$:

$$\begin{aligned} hyper(s) &= \{s' \mid hyper(s', s)\} \\ hyper_n(s) &= hyper(s) \cap Top_n \end{aligned}$$

El conjunt d'hiperònims d'un lema l, $hyper(l)$, és la unió dels hiperònims de cadascun dels seus synsets. També podem definir el conjunt d'hiperònims de nivell n o menor d'un lema l, $hyper_n(l)$:

$$\begin{aligned} hyper(l) &= \bigcup_{s \in sentits(l)} hyper(s) \\ hyper_n(l) &= hyper(l) \cap Top_n \end{aligned}$$

Finalment, definirem un kernel semàntic K_h entre dos lemes l i l' com la mida de la intersecció dels seus respectius conjunts d'hiperònims; i una família de kernels semàntics K_{hn} com la mida de la intersecció dels conjunts d'hiperònims de nivell n o menor:

$$\begin{aligned} K_h(l, l') &= |hyper(l) \cap hyper(l')| \\ K_{hn}(l, l') &= |hyper_n(l) \cap hyper_n(l')| \end{aligned}$$

Per últim, en comptes de considerar tots els sentits d'un lema l, podem considerar aquell que a WordNet es considera el més freqüent (essent la freqüència estimada a partir de corpus). Si anomenem aquest sentit més freqüent $smf(l) \in S$, podem definir conjunts d'hiperònims del sentit més freqüent, $hyper^f(l)$ i $hyper^f_n(l)$, i kernels que només tenen en compte el sentit més freqüent dels lemes, K^f_h i K^f_{hn}:

$$\begin{aligned} hyper^f(l) &= hyper(smf(l)) \\ hyper^f_n(l) &= hyper^f(l) \cap Top_n \\ K^f_h(l, l') &= |hyper^f(l) \cap hyper^f(l')| \\ K^f_{hn}(l, l') &= |hyper^f_n(l) \cap hyper^f_n(l')| \end{aligned}$$

Seguint la mateixa representació dels patrons que a la secció 4.1, de cara al clustering definim un conjunt de kernels semàntics K_*^S entre dos patrons ξ i ξ':

Tipus Usa només el tipus i el subtipus de les dues entitats relacionades:

$$K_T^S(\xi,\xi') = K^i(\tau_1,\tau_1') + K^i(\bar{\tau}_1,\bar{\tau}_1') + K^i(\tau_2,\tau_2') + K^i(\bar{\tau}_2,\bar{\tau}_2')$$

Morfològic Usa el tipus i la informació morfològica de les dues entitats relacionades. Equival al kernel K_A^Z de la secció 4.1

$$K_M^S(\xi,\xi') = K_{arg}(\xi,\xi)$$

Semàntic Usa el tipus i els hiperònims dels nuclis de les dues entitats relacionades. En podem definir diversos segons el conjunt d'hiperònims que considerem (tots els sentits o només el més freqüent, tots els nivells o només els n primers):

$$\begin{aligned}
K_S^S(\xi,\xi') &= K_T^S(\xi,\xi') + K_h(l_1,l_1') + K_h(l_2,l_2') \\
K_{SF}^S(\xi,\xi') &= K_T^S(\xi,\xi') + K_h^f(l_1,l_1') + K_h^f(l_2,l_2') \\
K_{Sn}^S(\xi,\xi') &= K_T^S(\xi,\xi') + K_{hn}(l_1,l_1') + K_{hn}(l_2,l_2') \\
K_{SFn}^S(\xi,\xi') &= K_T^S(\xi,\xi') + K_{hn}^f(l_1,l_1') + K_{hn}^f(l_2,l_2')
\end{aligned}$$

Morfo-Semàntic Usa la informació de tipus, morfològica i d'hiperonímia de les dues entitats. També se'n pot definir una família:

$$\begin{aligned}
K_{MS}^S(\xi,\xi') &= K_M^S(\xi,\xi') + K_h(l_1,l_1') + K_h(l_2,l_2') \\
K_{MSF}^S(\xi,\xi') &= K_M^S(\xi,\xi') + K_h^f(l_1,l_1') + K_h^f(l_2,l_2') \\
K_{MSn}^S(\xi,\xi') &= K_M^S(\xi,\xi') + K_{hn}(l_1,l_1') + K_{hn}(l_2,l_2') \\
K_{MSFn}^S(\xi,\xi') &= K_M^S(\xi,\xi') + K_{hn}^f(l_1,l_1') + K_{hn}^f(l_2,l_2')
\end{aligned}$$

5.2 Experiments

Per tal d'avaluar els nous kernels proposats, vam replicar els experiments de la secció 4.2, usant les mateixes dades, mètriques i condicions, però usant els kernels introduïts a l'apartat anterior.

Vam experimentar amb valors de $n \in \{1,2,3\}$ per al paràmetre n dels kernels K_{SFn}^S i K_{MSFn}^S, usant doncs fins als primers 3 nivells de la jerarquia de WordNet. Degut a la gran quantitat de característiques i, per tant, soroll que introdueix usar els synsets de tots els nivells de WordNet, vam optar per no utilitzar els kernels K_S^S, K_{SF}^S, K_{MS}^S i K_{MSF}^S en els experiments.

5.3 Resultats

Els resultats obtinguts poden veure's a la taula 6.

Pot observar-se com el kernel més senzill, K_T, que només té en compte la informació referent als tipus d'entitats relacionades, és el que obté millors resultats, millors fins i tot que els obtinguts per K_M^S, equivalent al K_A^Z que aconseguia les xifres més altes en la secció 4.3. Pel que fa als diversos kernels semàntics, en general l'adició de més nivells de la jerarquia fa empitjorar els resultats; i les versions que només tenen en compte la semàntica obtenen puntuacions més altes que les versions que també incorporen la morfologia. Respecte a si és millor usar tots els sentits de WordNet o només el més freqüent, no sembla haver-hi una tendència clara. Tanmateix, el segon millor resultat, després de K_T, l'obté el kernel K_{S1}^S, semàntic tenint en compte els synsets de nivell 1 de tots els sentits dels mots.

La tendència clara dels resultat a empitjorar a mida que s'afegeixen més nivells de la jerarquia sembla corroborar la intuïció que usar els sentits de tots els nivells de WordNet hauria introduït una quantitat excessiva de soroll al procés.

Tipus				Morfològic			
Kernel	Pur	IPur	F1	Prec	Rec	F1	Kernel
K_T	**0.658**	**0.527**	**0.586**	0.622	0.460	0.529	K_M^S
Semàntic (Tots)				Semàntic (Freqüent)			
Kernel	Pur	IPur	F1	Prec	Rec	F1	Kernel
K_{S1}^S	0.658	0.477	0.553	0.655	0.470	0.547	K_{SF1}^S
K_{S2}^S	0.600	0.453	0.516	0.632	0.479	0.545	K_{SF2}^S
K_{S3}^S	0.536	0.355	0.427	0.530	0.362	0.430	K_{SF3}^S
Morfo-Semàntic (Tots)				Morfo-Semàntic (Freqüent)			
Kernel	Pur	IPur	F1	Prec	Rec	F1	Kernel
K_{MS1}^S	0.622	0.417	0.499	0.616	0.407	0.490	K_{MSF1}^S
K_{MS2}^S	0.550	0.345	0.424	0.569	0.344	0.429	K_{MSF2}^S
K_{MS3}^S	0.521	0.315	0.393	0.527	0.332	0.407	K_{MSF3}^S

Taula 8: Resultats del clustering de patrons sobre el corpus ACE

5.4 Anàlisi

Creiem que una de les possibles raons per què els kernels semàntics no milloren els resultats del kernel més bàsic K_T^S rau en l'estratègia emprada per a retallar els synsets considerats de cada mot. El fet d'utilitzar només els nivells més alts de la jerarquia de WordNet fa que només es mantinguin synsets excessivament genèrics. Com a exemple, a la figura 10 es poden veure synsets de nivell 1, 2 i 3 de WordNet.

L'ús de conceptes tan genèrics fa que els conceptes clau, aquells que permeten distingir entre diversos tipus de relacions, quedin confosos amb molts altres d'irrellevants, representats per synsets d'un nivell massa alt. Sembla necessari utilitzar algun altre tipus d'estratègia, més guiada per les dades, per a determinar el nivell correcte a què tallar la jerarquia.

Amb aquesta intenció vam iniciar els darrers experiments de la meva estada.

Nivell	Synset
1	*entity*
2	*physical entity*
3	*object, physical object*
3	*causal agent, cause, causal agency*
3	...
2	*abstraction, abstract entity*
3	*psychological feature*
3	*attribute*
3	...
2	...
1	...

Figura 10: Exemples de synsets de nivell 1, 2 i 3 de WordNet

6 Clustering de Synsets de WordNet

Una aproximació per a reduir el nombre de synsets d'una forma guiada per les dades és utilitzar clustering. Identificant els sentits dels mots mitjançant clústers de synsets en comptes de synsets es redueix el nombre total de conceptes, agrupant els synsets relacionats en un únic concepte. Una bona estratègia de clustering hauria de permetre detectar el nombre de conceptes i el nivell de generalització o especialització necessari en cadascun d'ells.

Seguint amb l'aproximació que havíem estat usant durant tot aquest temps, vam optar per aplicar algorismes de clustering amb kernels, en concret kernel k-means, per a obtenir els clústers de synsets. Un cop obtinguts els clústers, van reformular en funció dels clústers de conceptes els kernels semàntics proposats a la secció 5.1.

6.1 Aproximació

Per tal de clusteritzar els sentits de WordNet, el primer pas és definir un kernel entre sentits. En el nostre cas, hem utilitzat la mesura de similitud semàntica de Resnik (1995). Aquesta mesura està basada en l'anomenat contingut d'informació. Donat un cert synset s, el seu contingut d'informació, $CI(s)$, és menys el logaritme de la seva probabilitat d'aparició, estimada a partir de corpus:

$$CI(s) = -\log p(s)$$

Donats dos synsets s i s', la similitud semàntica entre ells és el màxim dels contingut d'informació dels hiperònims comuns. Aquesta mesura és la que utilitzarem com a kernel entre synsets, K^R:

$$K^R(s, s') = \max\{CI(h) \mid h \in hyper(s) \cap hyper(s')\}$$

Per a evitar casos en què s i s' no tenen cap hiperònim comú, afegim una arrel comuna fictícia $s_\perp$ a la jerarquia de WordNet, que serà hiperònim de tots els synsets. Com que la probabilitat d'aparició d'aquest synset $s_\perp$ serà de 1, el seu contingut d'informació serà nul, i llavors:

$$hyper(s) \cap hyper(s') = \emptyset \longleftrightarrow K^R(s, s') = 0$$

Aquest kernel s'utilitza en un algorisme de clustering *on-line*, basat en l'algorisme de veïns més propers (*nearest neighbours*) de Lu i Fu (1978). L'adaptació que hem fet pot veure's a l'algorisme 1, i es basa en clústers en comptes d'en distàncies com la formulació original. El procediment és el següent:

1. Es defineix un valor llindar o radi de líder r.
2. Per a cada element:
   (a) Es troba el kernel entre ell i cadascun dels líders.
   (b) Si el kernel amb el líder més semblant és major que el radi r, l'element s'afegeix al clúster d'aquest líder.
   (c) Altrament, es crea un nou clúster amb l'element actual, que també s'afegeix a la llista de líders.

Els principals avantatges d'aquest algorisme són que és senzill, ràpid (només requereix una passada per la col·lecció) i detecta automàticament el nombre de clústers.

El procés comença per aplicar l'algorisme de veí més proper al conjunt dels sentits tots els mots que apareixen a la col·lecció de documents. El clustering resultant es pot denotar per Π_r^S. Es defineixen llavors una sèrie de kernels que, en comptes d'utilitzar tota la jerarquia de WordNet, només utilitzin, com a conceptes, els clústers de Π_r^S. Aquests clústers els denominarem *clústers-concepte* o *c-conceptes*.

Entrada: Conjunt de synsets S
Entrada: Kernel K_S
Entrada: Radi de líder r
Variable: Conjunt de líders Li
Sortida: Clustering Π

```
// Inicialitzar líder i clustering
Li = ∅;
Π = ∅;
per a cada s ∈ S fer
    // Buscar líder més proper
    max = 0;
    imax = ⊥;
    per a cada li_i ∈ Li fer
        // Si és més proper que el radi i que el màxim
        si K_S(s, li_i) > r ∧ K_S(s, li_i) > max llavors
            // Actualitzar el màxim
            max = K_S(s, li_i);
            imax = i;
        fiif
    fiper
    // Si hem trobat un màxim
    si imax ≠ ⊥ llavors
        // Afegir al clúster del màxim
        π_imax = π_imax ∪ {s};
    si no
        // Crear un nou clúster
        Π = Π ∪ {{s}};
        Li = Li ∪ {s};
    fiif
fiper
retorna Π;
```

Algorisme 1: Clustering de veí més proper amb kernels

Denotant els c-conceptes d'un cert mot, identificat pel seu lema l, com a $cconceptes_r(l)$:

$$cconceptes_r(l) = \bigcup_{s \in sentits(l)} \Pi_r^S(s)$$

es pot crear un kernel K_{Cr} basat en c-conceptes, com la mida de la intersecció dels conjunts de c-conceptes de dos mots:

$$K_{Cr}(l, l') = |cconceptes_r(l) \cap cconceptes_r(l')|$$

Una altra possibilitat és només utilitzar el c-concepte del seu sentit més freqüent, $cconcepte_r^f$, per a obtenir un kernel K_{Cr}^f, que només té en compte la identitat entre els c-conceptes del sentit més freqüent de dos mots:

$$\begin{aligned} cconcepte_r^f(l) &= \Pi_r^S(smf(l)) \\ K_{Cr}^f(l, l') &= K^i(cconcepte_r^f(l), cconcepte_r^f(l')) \end{aligned}$$

Usant aquests kernels entre mots, es poden definir versions dels kernels semàntics entre patrons presentats a la secció 5.1 usant c-conceptes en comptes de synsets:

$$\begin{aligned} K_r^C(\xi, \xi') &= K_T^S(\xi, \xi') + K_{Cr}(l_1, l_1') + K_{Cr}(l_2, l_2') \\ K_{Fr}^C(\xi, \xi') &= K_T^S(\xi, \xi') + K_{Cr}^f(l_1, l_1') + K_{Cr}^f(l_2, l_2') \end{aligned}$$

6.2 Experiments

De nou, per tal d'avaluar els nous kernels proposats, vam replicar els experiments de les seccions 4.2 i 5.2, usant les mateixes dades, mètriques i condicions, però amb els kernels introduïts a l'apartat anterior.

Com a valors del radi a l'algorisme de clustering de veí més proper per a l'obtenció dels c-conceptes, s'ha experimentat amb valors de $r \in \{1.0, 1.5, 2.0 \ldots 6.0\}$.

6.3 Resultats

Els resultats del clustering usant c-conceptes poden veure's a la taula 9. Només es mostren els resultats per a valors del radi a l'interval $r \in \{3.0 \ldots 5.0\}$, on se situen les execucions amb els millors resultats.

Com pot veure's, fer variar el valor del radi r no produeix grans variacions (tot just al voltant del 4%) en termes de cap de les mètriques. Els resultats tampoc no són concloents respecte a si és millor usar tots els synsets dels mots o només el més freqüent.

Tanmateix, el que també s'observa és que el coneixement que s'incorpora mitjançant els kernels basats en c-conceptes és menys sorollós que l'obtingut per les aproximacions semàntiques de la secció 5.1. Els valors de puresa són sempre més alts que usant usant el kernel basat només en tipus K_T^S. Tanmateix, la puresa inversa baixa, i els valors de F_1 acaben essent similars o lleugerament inferiors. El kernel $K_{3.5}^C$ és el que proporciona major F_1, obtenint el mateix valor que K_T^S.

Aquests resultats són interessants, i semblen corroborar que afegint semàntica de forma selectiva la qualitat dels kernels pot millorar.

6.4 Anàlisi

La taula 10 mostra algunes de les relacions que passen a ser ben classificades al canviar el kernel basat només en tipus, K_T^S, pel kernel basat en c-conceptes $K_{3.5}^C$; així com algunes que, pel contrari, K_T^S classificava bé i amb $K_{3.5}^C$ passen a confondre's amb altres categories. Molts dels casos es

Tipus				Morfològic			
Kernel	Pur	IPur	F1	Prec	Rec	F1	Kernel
K_T^S	0.658	**0.527**	**0.586**	0.622	0.460	0.529	K_M^S
C-Conceptes (Tots)				C-Conceptes (Freqüent)			
Kernel	Pur	IPur	F1	Prec	Rec	F1	Kernel
$K_{3.0}^C$	0.670	0.499	0.572	0.671	0.507	0.578	$K_{F3.0}^C$
$K_{3.5}^C$	0.672	0.520	**0.586**	0.668	0.488	0.564	$K_{F3.5}^C$
$K_{4.0}^C$	**0.684**	0.509	0.584	0.669	0.477	0.557	$K_{F4.0}^C$
$K_{4.5}^C$	0.677	0.513	0.584	0.675	0.484	0.564	$K_{F4.5}^C$
$K_{5.0}^C$	0.665	0.485	0.561	0.682	0.500	0.577	$K_{F5.0}^C$

Taula 9: Resultats del clustering de patrons sobre el corpus ACE

Millores	Nous Errors
You are ***my*** *best* ***friend*** Lasting $\not\rightarrow$ Family	***his brother*** Family $\rightarrow$ Lasting
I visited all ***their families*** Family $\not\rightarrow$ Business	***grandson*** *of the company's* ***founder*** Family $\rightarrow$ Business
a ***professor*** *of psychiatry at* ***NYU*** >Employment $\not\rightarrow$ >Student	***they*** *have a* ***daughter*** Family $\rightarrow$ Business
the ***secretary*** *of* ***state*** >Employment $\not\rightarrow$ >Founder	***team*** *in high* ***school*** >Subsidiary $\rightarrow$ >Student

Taula 10: Canvis amb l'ús de c-conceptes

poden comprendre a partir de la taula 11, que mostra alguns dels c-conceptes que es formen amb l'algorisme de veí més proper.

Així, a *You are my best friend*, la paraula clau de la relació, *friend*, es troba en un c-concepte amb mots com *affiliate, associate, colleague, fellow...*, que són clars indicadors d'una relació personal però de tipus no familiar. Al cas *I visited all their families*, és el mot *family* el que identifica el tipus de relació i el que es troba en un c-concepte amb altres mots com *clan, couple...*, que indiquen relacions familiars.

Tanmateix, el fet d'usar tots els sentits d'un mot també provoca errors. Així, existeixen sentits de *brother* i *sister* referents al seu ús per a indicar *amic* o *amiga*, fet que fa que apareguin al c-concepte dels mots referents a l'amistat com els ja mencionats *affiliate, associate, colleague, fellow...* i que relacions com *his brother* passin a classificar-se erròniament com a relació Lasting en comptes del tipus correcte Family.

El mètode presentat identifica doncs correctament conceptes que poden ser clau a l'hora de determinar el tipus d'una relació, però també afegeix errors, de forma que creiem que encara hi ha possibilitats per a millorar la nostra aproximació i, per tant, els resultats obtinguts.

7 Conclusions

Partint del propòsit inicial de millorar el clustering de documents usant patrons, durant la meva estada a la New York University vam dur a terme una sèrie d'experiments sobre kernels amb patrons, amb la idea de trobar una funció expressiva de similitud entre patrons.

Partint de kernels usats en tasques supervisades d'Extracció d'Informació, vam constatar que

clan, couple, family, folk, hanover, house, lancaster, line, name, people, side, tribe, windsor, york
aunt, babe, brother, child, cousin, father, mother, nephew, niece, offspring, parent, relative, sister, son, twin, uncle...
affiliate, associate, bride, brother, co-worker, colleague, fellow, friend, groom, mate, member, partner, peer, sister, teammate...
amazon, daughter, ex, exwife, female, gal, girl, lady, madam, mother, sister, sweetheart, widow, wife, woman...

Taula 11: Exemples de c-conceptes

els mètodes no supervisats poden patir una degradació important de rendiment amb l'addició d'informació no rellevant. Per a millorar la classificació de relacions en què el tipus ve expressat per la semàntica d'una paraula clau, vam intentar introduir informació semàntica als kernels. Finalment, mitjançant l'ús de tècniques de clustering també per a obtenir conceptes vam aconseguir kernels amb semàntica que identifiquen aquestes paraules clau.

Tanmateix, encara manca millorar els resultats d'aquesta aproximació, així com trobar una forma general d'integrar-hi més informació i distingir les característiques rellevants de forma no supervisada, per a dotar-la de major robustesa. Posteriorment, els kernels obtinguts per a clustering de patrons s'estendran i aplicaran per a clustering de documents a través dels patrons que contenen, dins la metodologia general per a obtenció no supervisada de patrons per a Extracció d'Informació que és l'objectiu de la meva tesi.

Agraïments

Gràcies a tot el projecte Proteus de la New York University i, en particular, al doctor Satoshi Sekine, que em va oferir l'oportunitat de realitzar una estada al seu grup.

Aquest treball ha estat realitzat amb el suport del Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya i del Fons Social Europeu.

A Notació

$\perp$	Valor indefinit
$\delta(\cdot,\cdot)$	Funció d'igualtat
$\lvert X\rvert$	Cardinalitat de X
$\lambda_i \in \Lambda$	Categories
$\pi_i \in \Pi$	Clústers
Pur	Puresa
$IPur$	Puresa inversa
F_1	Mètrica F_1
$d \in \mathcal{D}$	Document
$e \in E$	Dependència sintàctica
$f \in F$	Forma
$l \in L$	Lema
$t \in T$	Tag (Etiqueta sintàctica)
$\eta, \bar{\eta} \in H$	Esdeveniment
$\xi \in \Xi$	Patró
$\sigma \in \Sigma$	Funció sintàctica
$\omega \in \Omega$	Mot
$\nu \in N$	Entitat
$\tau \in T$	Tipus d'entitat
$\bar{\tau} \in \bar{T}$	Subtipus d'entitat
$s \in S$	Synset
Top_n	Synsets de nivell menor o igual a n
$sentits(\cdot)$	Sentits
$smf(\cdot)$	Sentit més freqüent
$hyperd(\cdot,\cdot)$	Hiperonímia directa
$hyper(\cdot,\cdot)$	Hiperonímia general
$hyper(\cdot)$	Hiperònims
$hyper_n(\cdot)$	Hiperònims de nivell menor o igual a n
$hyper^f(\cdot)$	Hiperònims del $smf(\cdot)$
$hyper_n^f(\cdot)$	Hiperònims del $smf(\cdot)$ de nivell menor o igual a n
$cconceptes_r(\cdot)$	C-conceptes
$cconcepte_r^f(\cdot)$	C-concepte del $smf(\cdot)$
$p(\cdot)$	Probabilitat
$\alpha, \vartheta \in \Theta$	Paràmetres de model
$\hat{\Theta}$	Estimació de paràmetres
$K(\cdot,\cdot)$	Kernel
$d(\cdot,\cdot)$	Distància
Φ	Espai vectorial
$\phi(\cdot)$	Funció de mapeig a espai vectorial

Referències

ACE05. The ACE 2005 (ACE05) evaluation plan, 2005.

A. Culotta i J. Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

G.F. DeJong. Prediction and substantiation: A new approach to natural language processing. *Cognitive Science*, 3, 1979.

A.P. Dempster, N.M. Laird, i D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society, Series B*, 39(1), 1977.

C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Y. Freund i R.E. Schapire. Large margin classification using the perceptron algorithm. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, 1998.

M. Girolami. Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, 13(3), 2002.

E. Gonzàlez. Una aproximació d'aprenentatge automàtic per a extracció d'informació adaptativa. Master's thesis, Departament de Llenguatges i Sistèmes Informàtics, Universitat Politècnica de Catalunya, 2007.

J. Hobbs. The generic information extraction system. In *Proceedings of the 5th Message Understanding Conference (MUC)*, 1993.

A.K. Jain, M.N. Murty, i P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3), 1999.

S. Lu i K.S. Fu. A sentence-to-sentence clustering procedure for pattern analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 8(5), 1978.

C. Manning i H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

J. McQueen. Some methods for classification and anlysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, 1909.

A. Meyers, R. Grishman, M. Kosaka, i S. Zhao. Covering treebanks with GLARF. In *Proceedings of the ACL Workshop on Human Language Technology and Knowledge Management*, 2001.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, i K. Miller. Five papers on WordNet. *International Journal of Lexicography*, 3(4), 1990. Special Issue.

K. Nigam, A. McCallum, S. Thrun, i T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2–3), 2000.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

E. Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI)*, 1993.

D. Roth i W. Yih. Relational learning via propositional algorithms: An information extraction case study. In *Proceedings of the 15th International Conference On Artificial Intelligence (IJCAI)*, 2001.

B. Scholkopf, A.J. Smola, i K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5), 1998.

M. Surdeanu, J. Turmo, i A. Ageno. A hybrid unsupervised approach for document clustering. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

M. Surdeanu, J. Turmo, i A. Ageno. A hybrid approach for the acquisition of information extraction patterns. In *Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM)*, 2006.

Jordi Turmo i Horacio Rodríguez. Learning rules for information extraction. *Natural Language Engineering*, 8(2/3), 2002.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

R. Xu i D. Wunsch, II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 2005.

R. Yangarber. Counter-training in discovery of semantic patterns. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

D. Zelenko, C. Aone, i A. Richardella. Kernel methods for relation extraction. *Journal of Machine Learning Research 3*, pages 1083–1106, 2003.

R. Zhang i A.I. Rudnicky. A large scale clustering scheme for kernel k-means. 4, 2002.

S. Zhao i R. Grishman. Extracting relations with integrated information using kernel methods. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

Y. Zhao i G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3), 2004.