Generació automàtica de resums de textos especialitzats: experimentacions en llengua catalana

Jorge Vivaldi, Iria da Cunha, Juan-Manuel Torres-Moreno i Patricia Velázquez-Morales
Universitat Pompeu Fabra i Universitat d'Avinyó

Jorge Vivaldi és doctor en informàtica per la Universitat Politècnica de Catalunya, investigador del grup IULATERM de l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra i especialista en corpus textuals i processament del llenguatge natural (extracció de terminologia).

Iria da Cunha és doctora en lingüística aplicada per la Universitat Pompeu Fabra, investigadora del grup IULATERM d'aquesta universitat i especialista en resum automàtic, anàlisi del discurs especialitzat i terminologia.

Juan-Manuel Torres-Moreno és doctor en informàtica per l'Institut National Polytechnique de Grenoble, investigador del grup TALNE del Laboratoire Informatique d'Avignon i especialista en resum automàtic, classificació automàtica de documents i aprenentatge automàtic.

Patricia Velázquez-Morales és doctora en ciència dels materials per l'Institut National Polytechnique de Grenoble, investigadora del VM Labs de la Universitat d'Avinyó i especialista en resum automàtic i constitució de corpus especialitzats multilingües.

Resum

En aquest article presentem un nou algoritme per al resum automàtic de textos especialitzats, que combina recursos terminològics (l'ús de l'extractor de termes YATE) i semàntics (consulta de l'ontologia lèxica EuroWordNet). Apliquem l'algoritme a un corpus de textos mèdics en català i n'avaluem els resums automàtics produïts, amb el sistema FRESA, tot comparant-los amb sengles resums baseline *i amb els resums d'un altre resumidor automàtic, el sistema OTS. L'algoritme proposat obté prou bons resultats, però el potencial de millora és, segons el nostre parer, molt alt.*

Paraules clau: *resum automàtic; textos especialitzats; català*

Abstract

In this article we present a new algorithm for the automatic summarisation of specialized texts, which combines terminological resources (YATE as a terminology extractor) and semantic resources (EuroWordNet as a lexical ontology). We apply this algorithm to a corpus of medical texts in Catalan and we evaluate the automatic summaries produced with the system FRESA, comparing them with baseline summaries and with results of another automatic summary system, the OTS. The new algorithm yields adequate results, but the potential for improvement is, in our view, very high.

Keywords: *automatic summarisation; specialized texts; Catalan*

Terminàlia 1 (2010): 26-32
DOI: 10.2436/20.2503.01.10 · ISSN: 2013-6692

Generació automàtica de resums de textos especialitzats: experimentacions en llengua catalana

Jorge Vivaldi, Iria da Cunha, Juan-Manuel Torres-Moreno i Patricia Velázquez-Morales

1 Introducció

Un resum és una representació abreujada d'un text original. L'objectiu principal del resum és que el lector pugui conèixer els punts essencials del text original sense la necessitat de llegir el text sencer i que, per tant, pugui decidir més ràpidament si el text original conté o no la informació que està buscant. Per aquest motiu, avui dia el resum automàtic és un tema de recerca molt rellevant. Durant els anys seixanta, la recerca en aquest camp se centrava, bàsicament, en el discurs general, encara que hi havia algunes excepcions, com els primers experiments amb textos tècnics de Luhn (1959) i, més endavant, el resumidor automàtic de textos químics de Pollock i Zamora (1975). Durant els anys noranta, uns quants investigadors van començar a treballar en el resum del discurs especialitzat (per exemple Paice, 1990; Riloff, 1993; Lehmam, 1995; McKeown i Radev, 1995; Abracos i Lopes, 1997) i l'any 2000, Saggion i Lapalme; però per norma utilitzaven les mateixes estratègies que les que s'empraven per al discurs general: estructura textual o del discurs, frases clau, posició de frase, entitats denominades, tècniques estadístiques, aprenentatge automàtic, etc. Afantenos *et al.* (2005) assenyalen que, en concret, el resum automàtic de textos mèdics ha esdevingut un tema de recerca crucial perquè els professionals de l'àmbit biomèdic necessiten processar una gran quantitat de documents. Hi ha alguns treballs pertinents relacionats amb aquest camp, com per exemple Damianos *et al.* (2002), Johnson *et al.* (2002), Gaizauskas *et al.* (2001), Lenci *et al.* (2002) i Kan *et al.* (2001); de nou, però, les tècniques utilitzades no són específiques del domini.

Un treball que sí que té en compte recursos específics del domini temàtic és el de Reeve *et al.* (2007). Aquest sistema utilitza cadenes lèxiques, és a dir, seqüències de paraules entre les quals hi ha una relació lexicosemàntica (principalment d'identitat o de sinonímia). Però cal dir que en aquest cas es limita als sintagmes nominals que tenen associat un tipus semàntic a l'UMLS.[1]

Pel que fa a l'àmbit específic de la llengua catalana l'únic treball de què tenim coneixement és el que fa referència al resum automàtic de notícies de premsa de Fuentes *et al.* (2004), que se centra en el discurs general periodístic i no en textos especialitzats.

Inspirats en el treball de Luhn (1959), que establia que els termes que apareixen en el títol d'un text científic són marques pertinents que n'indiquen el tema principal, i també pels treballs de Barzilay i Elhadad (1997) i de Silber i McCoy (2000), que utilitzaven cadenes lèxiques amb relacions lexicosemàntiques, hem dissenyat una estratègia nova de resum per a textos especialitzats. La hipòtesi és que els termes que es relacionen semànticament amb els termes inclosos en el títol d'un text especialitzat són especialment pertinents per al resum. Així, considerem que un algoritme de resum automàtic que seleccioni les frases del text que inclouen no només els termes del títol, sinó també les frases que tenen termes que es relacionen semànticament amb els anteriors, aconseguirà resultats positius. Pel que sabem, no existeixen sistemes de resum basats en aquesta combinació de recursos terminològics i semàntics. Per tant, el nostre treball contribueix a la creació d'una línia nova de recerca en resum automàtic.

Per a donar suport a aquesta idea, hem dissenyat un nou algoritme de resum basat en aquest principi, l'hem aplicat sobre un corpus de textos mèdics en català i hem avaluat els resultats utilitzant el FRESA (*Framework of Evaluation of Summaries Automatically*) (Torres-Moreno, 2010; Torres-Moreno *et al.*, en premsa). En l'apartat 2, descrivim la metodologia de la nostra tasca; a continuació, en l'apartat 3, descrivim l'algoritme de resum; en l'apartat 4 presentem els experiments fets i els resultats corresponents; i finalment en l'apartat 5 mostrem les conclusions del nostre treball.

2 Recursos utilitzats

Com a primer pas del nostre treball, hem dissenyat l'algoritme de resum. Després, hem compilat una col·lecció de vint textos que provenen del subcorpus de medicina del Corpus Tècnic de l'Institut Universitari de Lingüística Aplicada (IULA) (Vivaldi, 2009). A continuació, hem aplicat el nostre algoritme de resum, i finalment els resultats s'han avaluat utilitzant el FRESA.

Com es mostrarà en l'apartat 3, l'algoritme de resum es basa en la utilització de l'extractor de candidats a terme YATE (*Yet Another Term Extractor*) (Vivaldi, 2001; Vivaldi i Rodríguez, 2001*a*, 2001*b*). YATE és un extractor de termes que presenta les següents característiques principals:

a) Utilització intensiva d'informació semàntica (obtinguda a través de l'EuroWordNet (EWN).[2]
b) Utilització d'una metodologia híbrida que combina diferents estratègies d'extracció de candidats a terme (lingüístiques i estadístiques).

Inicialment, el YATE es va desenvolupar per a l'àmbit mèdic, encara que s'ha adaptat a altres dominis, com ara la genètica i l'economia, i ara s'està adaptant a d'altres.

La metodologia habitual per a avaluar resums és la utilització del ROUGE (*Recall-Oriented Understanding for Gisting Evaluation*) (Lin, 2004). Aquest sistema compara la coaparició de *n*-grames en el resum produït per la màquina amb un o més resums model o de referència. Aquests models de resum normalment els fa l'autor o altres especialistes del domini, però també poden ser resultat d'aplicar altres metodologies de resum. En qualsevol cas, la utilització del ROUGE requereix disposar, com a mínim, d'un resum manual. En el cas dels textos del Corpus Tècnic de l'IULA no disposem dels resums respectius redactats pels autors, com

seria habitual en els articles mèdics de revistes especialitzades d'aquest àmbit, per exemple. Aconseguir resums d'aquests textos elaborats per persones seria molt costós en temps i diners, de manera que poder comptar amb un sistema d'avaluació de resums que no necessiti resums de referència és un gran avantatge. Per aquest motiu hem utilitzat el FRESA per a avaluar els resums que fa el nostre algoritme, en comptes del ROUGE.

El FRESA és un sistema per avaluar resums automàtics sense necessitat de comptar amb resums model o de referència redactats per humans. Utilitza una divergència de probabilitats per a comparar la distribució de probabilitats dels *n*-grames del text original i del resum produït. La idea d'emprar les divergències de probabilitats per a avaluar resums prové dels treballs de Lin *et al.* (2006) i de Louis i Nenkova (2008, 2009). Tanmateix, Lin *et al.* (2006) utilitzen la divergència de Küllback-Leibler, que presenta alguns problemes —com l'asimetria— que en limiten la utilitat pràctica. Els treballs de Louis i Nenkova (2008, 2009) es limiten a l'estudi de la divergència de Jensen-Shanon (JS) de la distribució dels unigrames, mentre que el FRESA calcula la divergència JS de la distribució de probabilitat dels unigrames, dels bigrames, dels bigrames amb buits (tal com fa el ROUGE SU-4)[3] i la mitjana de les tres distribucions. A més, detecta la llengua del text estadísticament (castellà, català, francès i anglès) i realitza un processament lingüístic superficial, eliminant les paraules funcionals de cada llengua, abans de calcular les distribucions de probabilitats.

3 Disseny de l'algoritme

La idea general d'aquest algoritme de resum és obtenir un resultat de pertinència per a cada frase, tenint en compte tant la terminologicitat[4] dels termes trobats en les frases com la similitud entre aquests termes i els trobats en el títol del document. Inicialment, l'extractor YATE s'utilitza per a trobar els termes en els textos mèdics que cal resumir. Cada terme tindrà una terminologicitat associada. Internament, l'extractor identifica els termes presents en el títol i els presents en el cos del document. Llavors, un mòdul de YATE mesura la distància semàntica entre cadascun dels termes presents en el cos de l'article i tots els termes trobats en el títol. Com a resultat, el terme rep un coeficient que es calcula utilitzant [1]:

$$Score(t) = T(t)P + MaxSim(t{:}t_t)(1 - P) \qquad [1]$$

en què t és una unitat terminològica, $T(t)$ és la seva terminologicitat, t_t és qualsevol terme que pertany al títol del document i P és un coeficient de ponderació (0,5 per defecte). La idea és que el resultat associat amb cada terme tindrà en compte tant la terminologicitat com la seva relació amb el títol del document. El valor per defecte assigna la mateixa rellevància tant a la terminologicitat com a la relació amb els termes del títol.

Per calcular la similitud entre els termes trobats en el títol i els presents en el cos del document, utilitzem informació obtinguda mitjançant els camins d'hiperonímia per a cada *synset* d'EuroWordNet (EWN). Per aquest propòsit utilitzem la fórmula [2]:

$$Sim(synset_1, synset_2) = \frac{2 \times CommonNodes(synset_1, synset_2)}{Depth(synset_1) + Depth(synset_2)} \qquad [2]$$

En la pràctica, la similitud entre dos termes mèdics com *vas* i *glàndula* es calcula com es mostra en la figura 1:

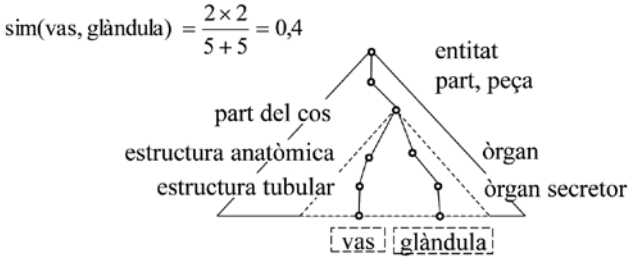

Figura 1. Exemple de càlcul de la similitud

En el cas dels termes complexos, s'analitzen tots els components (noms i adjectius), però només s'escull el component que ofereix la màxima similitud, amb l'objectiu d'explotar també la similitud entre els components del terme. En el cas dels adjectius, només s'utilitzen els que siguin relacionals, i la relació semàntica es compta utilitzant el nom corresponent (*bronqui*, *bronquial*). Per obtenir el resultat final FS*i* de cada frase (per $i = 1, \ldots, k$; en què k = nombre de frases del document), tenim en compte el resultat de tots els termes que cada frase inclou, utilitzant la fórmula següent [3]:

$$FS(s) = \sum_{t \in L_{CAT}(s)} Score(t) \qquad [3]$$

En què s és una frase del cos del document, $L_{CAT}(s)$ són els termes detectats en s i t és un terme. Per exemple, imaginem un article mèdic amb el títol següent: «Visites inapropiades al servei d'urgències d'un hospital general». Una de les frases d'aquest text podria ser: «Aquest és un estudi descriptiu d'una mostra aleatòria de 84.329 pacients visitats el 1999». Aquesta frase no inclou cap terme del títol, però conté el terme *pacient*, que a l'EWN està semànticament relacionat amb *hospital general* (un terme del títol). Específicament, la similitud entre aquests dos termes és 0,15. A més a més, el YATE assigna al terme *pacient* una terminologicitat d'1. En aquest cas i considerant $P = 0,5$, el resultat d'aquesta frase seria FS(s) = 0,15 x 0,5 + 1 x 0,5 = 0,575. Aquesta similitud té sentit perquè els pacients són usuaris

dels hospitals i per aquest motiu és lícit pensar que existeix una relació entre els dos passatges.

Per elaborar el resum final obtenim el resultat de totes les frases del text i les classifiquem d'acord amb el resultat. Després de decidir el nombre de frases que ha d'incloure el resum, escollim les frases que tenen el resultat més alt i les tornem a posar en l'ordre original.

4 Experiments i resultats

Per avaluar l'algoritme, l'apliquem sobre un corpus de vint textos del subcorpus de medicina del Corpus Tècnic de l'IULA. Els temes dels textos són variats: malaltia d'Alzheimer, glucosa, genoma humà, síndrome de Down, clonació humana, càncer, etc. El nombre total de paraules d'aquests textos és al voltant de 60.000 (el més extens consta de 6.046 paraules i el més curt n'inclou 1.183). Tenint en compte que els textos no tenen la mateixa extensió (encara que es manté dins d'uns límits), decidim extreure per als respectius resums un 20 % de les frases de cada text.

Com ja s'ha mencionat en l'apartat 2, avaluem els resums produïts pel nostre algoritme utilitzant el sistema FRESA. Perquè l'avaluació fos més completa, s'haurien de comparar els resultats obtinguts amb els resultats d'altres sistemes de resum. Tanmateix, no hi ha gaires sistemes de resum automàtics disponibles per al català. Així doncs, hem fet dos tipus de resums *baseline*. D'una banda, hem fet resums dels vint textos del nostre corpus que inclouen un 20 % de frases de cada text seleccionades aleatòriament (*baseline_ale*) i, de l'altra, hem fet resums que inclouen les frases que suposen el 20 % inicial del text (*baseline_prim*). També hem utilitzat per a la comparació resums del sistema Open Text Summarizer (OTS).[5] Aquest sistema és multilingüe i gratuït i, a més a més, permet resumir textos en català, tot i que cal especificar que els recursos de què disposa per al català són molt escassos, la qual cosa fa empitjorar-ne els resultats.

La taula 1 inclou els resultats del FRESA-M, és a dir, la mitjana de les tres distribucions de probabilitat (unigrames, bigrames i bigrames amb buits). La primera columna mostra l'identificador de cada text i les següents columnes mostren els resultats obtinguts pel nostre algoritme, la *baseline_prim*, la *baseline_ale* i el sistema OTS. S'ha destacat en negreta el millor resultat per a cada text. La figura 2 reflecteix de forma gràfica els resultats obtinguts.

Com es pot apreciar en la taula 1, els resums del nostre algoritme són els que, en general, obtenen els millors resultats (només superats per la *baseline_ale* en el cas dels textos m00163 i m00515, i per la *baseline_prim* en el cas del text m00165). La figura 3 mostra els resultats del nostre algoritme amb el FRESA-1, FRESA-2, FRESA-SU4 i FRESA-M. Observem que els millors resultats són els obtinguts amb l'avaluació del FRESA-1, cosa lògica ja que aquesta mesura només té en compte els unigrames.

Taula 1. Resultats del FRESA-M (mitjana del FRESA-1, FRESA-2 i FRESA-SU4)

Text	Algoritme	*Baseline_prim*	*Baseline_ale*	OTS
m00071	**0,85580**	0,77830	0,84678	0,84533
m00087	**0,85311**	0,76840	0,83193	0,83202
m00163	0,82005	0,83636	**0,86824**	0,86045
m00165	0,88127	**0,88300**	0,84741	0,82052
m00286	**0,87936**	0,82551	0,83937	0,86215
m00309	**0,86326**	0,81959	0,83586	0,85331
m00339	**0,89673**	0,83670	0,81551	0,84860
m00345	**0,84942**	0,79083	0,83669	0,84286
m00474	**0,85094**	0,80416	0,84039	0,83210
m00475	**0,86333**	0,83811	0,83781	0,85623
m00514	**0,87852**	0,81473	0,84999	0,84483
m00516	**0,86362**	0,84178	0,84328	0,84006
m00519	0,84782	0,81966	**0,85041**	0,81681
m00550	**0,85558**	0,79773	0,82557	0,84011
m00562	**0,85741**	0,78899	0,84597	0,82417
m00566	**0,85283**	0,79551	0,84690	0,82746
m00611	**0,85799**	0,81830	0,84253	0,84561
m00613	**0,87663**	0,84059	0,86351	0,84522
m00630	**0,88803**	0,79889	0,87117	0,84204
m00635	**0,84987**	0,80732	0,84277	0,84045
Mitjana	**0,86208**	0,81522	0,84410	0,84102

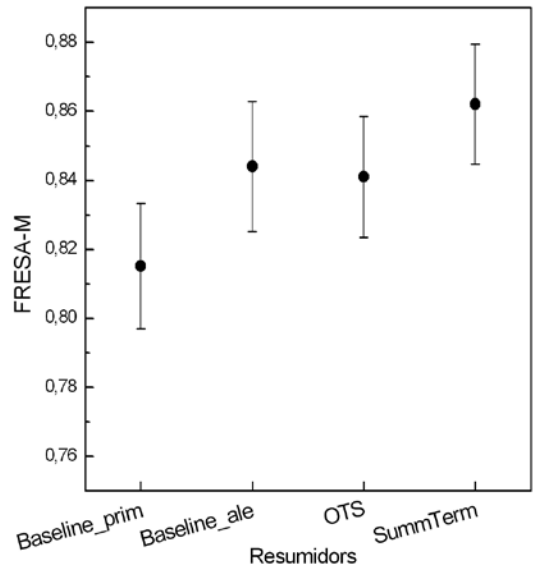

FIGURA 2. Resultats del FRESA-M expressats gràficament

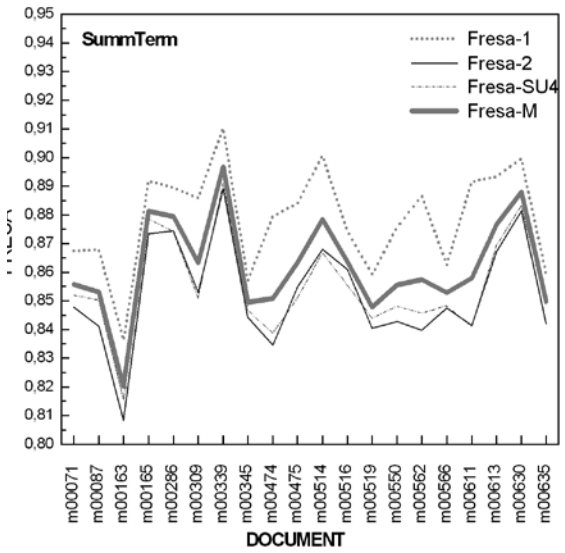

FIGURA 3. Resultats de l'algoritme amb el FRESA-1, FRESA-2, FRESA-SU4 i FRESA-M

Una anàlisi detallada dels resultats mostra que YATE és massa conservador, és a dir, proporciona com a termes només les seqüències que tenen una alta fiabilitat, sobretot en el cas dels termes multiparaula. Avancem que potser, atès l'alt nivell d'especialització dels documents del corpus, YATE hauria de relaxar les seves consideracions sobre què és un bon candidat de terme. Futurs experiments aniran per aquesta línia.

La llista de candidats a termes també mostra que els termes d'aquest tipus de textos són seqüències normalment complexes (com ara un nom seguit de dos o més adjectius) que apareixen més freqüentment del que es considera habitual (*sistema nerviós central*, *lesions tumorals papil·lars superficials*, *malformacions cardíaques congènites*, *oftalmoplegia externa progressiva*, etc.).

Un altre problema, específic del català, és que un o més components d'un candidat a terme no formin part de les entrades de l'EWN. En aquest sentit, hem trobat que formes relativament comunes de noms i adjectius, com ara *trombosi*, *alteració*, *cèl·lula*, *farmacèutic*, *sanguini*, *gènic*, entre d'altres, no estan incloses a l'EWN, cosa que dificulta la tasca de YATE. Un altre problema és l'aparició de paraules constituïdes per formants cultes que no es poden descompondre a causa de la seva complexitat imprevista.

També cal notar que la llargada mitjana dels textos és de 2.900 paraules. Aquesta llargada és prou bona per a fer un resum automàtic, però és massa curta per a permetre a YATE de beneficiar-se dels mètodes estadístics que incorpora. Millores en la cobertura de l'EWN, així com en el tractament de combinacions de formants cultes amb paraules regulars, haurien d'ajudar a millorar el rendiment del conjunt.

5 Conclusions

La conclusió principal del nostre treball és que l'algoritme de resum automàtic que hem dissenyat és vàlid per a textos en català. Avui dia no hi ha resumidors específics per a aquesta llengua, i encara menys resumidors de textos especialitzats. Creiem que l'algoritme de resum automàtic presentat és molt innovador, ja que combina recursos terminològics i semàntics. Els resultats obtinguts són bons, comparats amb els de les *baselines* i els de l'altre sistema de resum automàtic, l'OTS. De tota manera, creiem que és necessari continuar treballant en aquesta línia de recerca per tal de millorar-ne els resultats. També caldria introduir algunes millores a l'extractor YATE. Preveiem fer més experiments, canviant, per exemple, el factor de pes estadístic, donant més importància a la terminologicitat o a la similitud semàntica; i també volem avaluar resums d'altres mides.

Bibliografia

Abracos, José; Lopes, Gabriel (1997). «Statistical methods for retrieving most significant paragraphs in newspaper articles». A: *Intelligent scalable text summarization*. Madrid: Universidad Nacional de Educación a Distancia, p. 51-57.

Afantenos, Stergos; Karkaletsis, Vangelis; Stamatopoulos, Panagiotis (2005). «Summarization of medical documents: A survey». *Artificial Intelligence in Medicine*, vol. 33, núm. 2, p. 157-177.

Barzilay, Regina; Elhahad, Michael (1997). «Using lexical chains for text summarization». A: *Workshop on Intelligent scalable text summarization*. Madrid: Association for Computacional Linguistics, p. 10-17.

Damianos, Laurie [*et al.*] (2002). «Real users, real data, real problems: The MiTAP system for monitoring bio events». A: *Conference on Unified Science & Technology for Reducing Biological Threats & Countering Terrorism (BTR 2002)*. Mèxic, p. 167-177.

Fuentes, María; Gonzàlez, Edgar; Rodríguez, Horacio (2004). «Resumidor de notícies en català del projecte Hermes». II Congrés d'Enginyeria en Llengua Catalana (Andorra).

Gaizauskas, Robert [*et al.*] (2001). «Intelligent access to text: Integrating information extraction technology into text browsers». A: *Human Language Technology Conference*. San Diego (CA): Association for Computacional Linguistics, p. 189-193.

Johnson, David [*et al.*] (2002). «Modeling medical content for automated summarization». A: *Annals of the New York Academy of Sciences*, núm. 980, p. 247-258.

Kageura, Kio; Umino, Bin (1996). «Methods of automatic term recognition: A review». *Terminology*, vol. 3, núm 2, p. 259-289.

Kan, Min-Yen; McKeown, Kathleen; Klavans, Judith (2001). «Domain-specific informative and indicative summarization for information retrieval». A: *Text summarization. Document Understanding Conference*. Nova Orleans (LA), p. 19-26.

Lehmam, Abderrafih (1995). *Le résumé des textes techniques et scientifiques, aspects linguistiques et computationnels*. Tesi de doctorat. Universitat Nancy 2.

Lenci, Alessandro [*et al.*] (2002). «Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project». A: *Third International Conference on Language Resources and Evaluation* (Las Palmas, 2002), París: European Resources Association, p. 1464-1471.

Lin, Chin-Lew (2004). «ROUGE: A package for automatic evaluation of summaries». A: *Text summarization branches out: Proceedings of the ACL-04 Workshop*. Barcelona: Association for Computational Linguistics, p. 25-26

Lin, Chin-Yew [*et al.*] (2006). «An information-theoretic approach to automatic evaluation of summaries». A: *Conference of the North American Chapter of the Association of Computational Linguistics*. Morristown (NJ): Association for Computacional Linguistics, p. 463-470.

Louis, Annie; Nenkova, Ani (2008). «Automatic summary evaluation without human models». A: *SIGIR'2008, Text Analysis Conference (TAC)*. Singapur: ACM Press, p. 306-314.

Louis, Annie; Nenkova, Ani (2009). «Automatically evaluating content selection in summarization without human models». A: *Conference on Empirical Methods in Natural Language Processing*. Singapur: Association for Computacional Linguistics.

Luhn, Hans (1959). «The automatic creation of literature abstracts». *IBM Journal of Research and Development*, núm. 2, p. 159-165.

McKeown, Kathleen; Radev, Dragomir (1995). «Generating summaries of multiple news articles». A: *SIGIR'95, 18th Annual Internacional Conference on Research and Development in Information Retrieval*. Seattle (WA): ACM Press, p. 74-82.

Paice, Chris (1990). «Constructing literature abstracts by computer: Techniques and prospects». *Information Processing and Management*, núm. 26, p. 171-186.

Pollock, Joseph; Zamora, Antonio (1975). «Automatic abstracting research at the chemical abstracts service». *Journal of Chemical Information and Computer Sciences*, vol. 15, núm. 4, p. 226-232.

Reeve, Lawrence; Han, Hyoil; Brooks, Ari (2007). «The use of domain-specific concepts in biomedical text summarization». *Information Processing and Management*, vol. 43, núm. 6, p. 1765-1776.

Riloff, Ellen (1993). «A corpus-based approach to domain-specific text summarisation: A proposal». A: Endres-Niggemeyer, Brigitte; Hobbs, Jerry; Sparck-Jones, Karen (ed.). *Workshop on summarising text for intelligent communication - Dagstuhl seminar report 79*. Dagstuhl (Alemanya).

Saggion, Horacio; Lapalme, Guy (2000). «Concept identification and presentation in the context of technical text summarization». A: *NAACL-ANLP Workshop on Automatic Summarization*. Seattle (WA): Association for Computacional Linguistics.

Silber, Gregory; McCoy, Kathleen (2000). «Efficient text summarization using lexical chains». A: *International Conference on Intelligent User Interfaces*. Nova York: ACM Press, p. 252-255.

Torres-Moreno, Juan-Manuel (2011 [2010]). «FRESA 1.0 (a FRamework for Evaluating Summaries Automatically)» [en línia]. Report tècnic LIA-RT-3-2010, Avinyó. <http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/fich_art/fresa.pdf> [Consulta: 9 juny 2010].

Torres-Moreno, Juan-Manuel [*et al.*] (en premsa). «Évaluation automatique de résumés avec et sans référence». A: *17e Conférence sur le Traitement Automatique des Langues Naturelles*. Mont-real (Canadà): Université de Montréal. École Polytechnique de Montréal.

Vivaldi, Jorge (2001). *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*. Tesi doctoral. Universitat Politècnica de Catalunya.

Vivaldi, Jorge (2009). «Corpus and exploitation tool: IULACT and bwanaNet». A: Cantos Gómez, Pascual; Sánchez Pérez, Aquilino (ed.). *A survey on corpus-based research*. Múrcia: Asociación Española de Lingüística de Corpus, p. 224-239.

Vivaldi, Jorge; Rodríguez, Horacio (2001*a*). «Improving term extraction by combining different techniques». *Terminology*, vol. 7, núm. 1, p. 31-47.

Vivaldi, Jorge; Rodríguez, Horacio (2001*b*). «Improving term extraction by system combination using boosting». *Lecture Notes in Computer Science*, núm. 2167, p. 515-526.

Vossen, Piek. (2004). «EuroWordNet: a multilingual database of autonomous and language specific wordnets connected via an Inter-Lingual-Index». *International Journal of Lexicography*, vol. 17, núm. 2, p. 161-173.

Notes

1. L'UMLS (Unified Medical Language System) és un tesaurus molt usat de l'àmbit biomèdic que proporciona una representació del coneixement en aquest domini. Aquesta representació consisteix en la classificació dels conceptes per tipus semàntics i de les relacions (jeràrquiques i no jeràrquiques) existents entre els tipus: vegeu *http://www.nlm.nih.gov*.
2. EuroWordNet (Vossen, 2004) és una base de dades lèxiques multilingüe de propòsit general, basada en WordNet, que inclou tant el català com altres llengües europees. S'estructura en unitats lexicosemàntiques (o *synsets*) enllaçades mitjançant relacions semàntiques bàsiques.
3. Un bigrama, en aquest context, és qualsevol seqüència de dues paraules. De vegades, aquestes paraules poden ser no consecutives. Aquest és el cas del ROUGE SU-4 en què hi pot haver un màxim de quatre paraules entre les dues que formen el bigrama.
4. La *terminologicitat* (de l'anglès *termhood*) ha estat definida per Kageura *et al.* (1996) com el grau de pertinència a un cert domini d'un candidat a terme.
5. Vegeu *http://libots.sourceforge.net*.