Jaume Martí i Marina Salse (coord.)
La terminologia i la documentació: relacions i sinergies
Barcelona: Institut d'Estudis Catalans, 2010, p. 75-83 DOI: 10.2436/15.2503.02.3

Ús d'estratègies estadístiques per a l'extracció automàtica d'unitats terminològiques

Mercè Vázquez, Antoni Oliver
Universitat Oberta de Catalunya
Barcelona

Resum

La detecció automàtica d'unitats lèxiques de caràcter especialitzat d'un determinat àmbit de coneixement és un dels reptes clau per a l'organització i la recuperació d'informació. En aquesta comunicació es planteja l'ús de diferents estratègies estadístiques, amb l'objectiu de poder extreure automàticament unitats terminològiques d'un àmbit d'especialitat a fi de recuperar i organitzar la informació que conté.

Paraules clau: classificació de documents, documentació, extracció de terminologia, mètodes estadístics, ontologia, recuperació d'informació, taxonomia.

Abstract: ***The use of statistics-based strategies for the automatic extraction of terminology units***

The automatic detection of lexical units of a specialised nature in a given area of knowledge is one of the key challenges in the organisation and retrieval of information. This communication addresses the use of different statistics strategies with a view to be able to automatically extract terminological units from a specialist area to retrieve and organise the information it contains.

Key words: classification of documents, documentation, terminology extraction, statistical methods, ontology, information retrieval, taxonomy.

1. Terminologia i documentació

Els àmbits de coneixement de la terminologia i la documentació s'han centrat, d'una banda, en la identificació i compilació dels termes i, de l'altra, en la identificació i compilació dels documents. Aquestes dues branques de coneixement han anat avançant en la descoberta de noves tècniques per a millorar llurs

processos de treball, però han tingut poques ocasions de compartir l'expertesa assolida en cada una de les àrees. En els darrers anys s'ha vist la necessitat de començar a compartir coneixement i establir lligams entre els especialistes d'aquestes dues àrees per a assolir resultats que puguin ser aprofitats en les dues àrees de coneixement.

En l'àmbit de la terminologia, el reconeixement automàtic d'unitats terminològiques i la detecció precoç de neologismes són alguns dels reptes que encara té pendents actualment el treball terminològic, els quals constitueixen la base de la proposta que es fa en el present article. La tasca de detecció automàtica d'unitats terminològiques i la compilació d'aquestes unitats permet disposar de material terminològic actualitzat, cada vegada més necessari per l'augment exponencial de recursos digitals, el problema d'accés als continguts i la dificultat que hi ha en l'automatització del contingut dels corpus. En aquest sentit, l'àmbit de la documentació necessita tenir a l'abast recursos terminològics que puguin explotar grans volums de corpus per a poder-ne extreure llistes de paraules clau, útils per a la indexació de continguts; elaborar taxonomies i, en última instància, crear ontologies. Així, doncs, s'estableix un marc d'interacció de coneixement i aprofitament de recursos molt important.

D'altra banda, la introducció d'estratègies estadístiques en el procés d'identificació d'unitats candidates a ser termes fa possible de treballar amb corpus d'especialitat de gran volum que poden ser monolingües, bilingües o multilingües i recuperar els equivalents corresponents de traducció i els contextos d'ús. Els mètodes estadístics reconeixen les unitats terminològiques a partir de la freqüència que tenen en un corpus marcat temàticament. Malgrat ser un càlcul molt senzill, el problema que presenta és que es fa difícil de recuperar termes que apareixen poques vegades en un corpus d'especialitat; per aquest motiu, s'ha de combinar amb l'ús de mesures estadístiques. Així, si es compara el valor de freqüència que té una unitat dins un corpus d'especialitat amb els resultats que ofereixen un conjunt de mesures estadístiques hi ha una evidència superior del caràcter terminològic d'un candidat a terme, ja que mesuren el nivell o grau d'associació de les unitats que constitueixen un candidat a terme.

2. Estratègia d'identificació d'unitats especialitzades

La tria d'una mesura estadística que sigui adequada per a identificar el major nombre de termes d'un corpus d'especialitat segueix un procés de preparació de resultats que comença amb l'extracció automàtica de candidats a terme del corpus i el filtratge d'aquests candidats amb una llista de paraules buides (conjuncions, preposicions, locucions, etc.), a fi de disposar d'una llista de candidats endreçats per freqüència.

clear forward signal 442 903 710 4358 540 671 455
data link layer 322 1589 1554 464 564 322 334
coast earth station 256 279 1007 954 274 256 626
earth station antenna 81 961 1150 279 677 85 97
earth station equipment 51 961 1150 1648 677 57 58
earth station antennas 29 961 1150 149 677 29 37
earth station Hpa 29 961 1150 104 677 29 32
earth station receiver 24 961 1150 134 677 24 35
earth station transmit 21 961 1150 81 677 24 21
earth station identification 16 961 1150 291 677 16 33
earth station HPàs 14 961 1150 35 677 16 16
earth station receive 13 961 1150 84 677 13 13
earth station complexes 12 961 1150 12 677 12 12
earth station located 10 961 1150 134 677 19 10
earth station transmitter 10 961 1150 42 677 10 10
earth station owner 8 961 1150 8 677 8 8
earth station number 2 961 1150 994 677 2 42

FIGURA 1. Llista de candidats a terme ordenats per valor de freqüència

En la imatge superior (figura 1) observem una mostra de candidats a terme filtrats amb una llista de paraules buides i endreçats per freqüència que pertanyen a un corpus d'especialitat de l'àmbit de les telecomunicacions. El valor de freqüència correspon al primer valor que apareix al costat del candidat a terme. La resta de valors corresponen al nombre de vegades que apareixen juntes en el corpus les diferents paraules que formen el candidat a terme. Així, el candidat «clear forward signal» veiem que apareix 442 vegades en el corpus, i així successivament.

A partir d'aquí, aquest resultat és processat amb tretze mesures estadístiques[1] que permeten de calcular una puntuació i un valor de rang per a cada candidat i mostren el resultat obtingut en ordre ascendent. La puntuació que s'atribueix a cada candidat indica si hi ha evidència o no n'hi ha que pugui ser una unitat terminològica.

A partir de la informació de freqüència i la puntuació que s'obté per a cada candidat a terme s'observa en quina posició queda endreçat i també quin valor de rang queda atribuït a cada candidat, tenint en compte que els candidats que tenen una mateixa puntuació queden aglutinats dins un mateix valor de rang. D'aquesta manera, els candidats que tenen un valor de rang més baix i una puntuació més alta corresponen a combinacions poc habituals i, per tant, hi ha una probabilitat més alta que siguin terminològiques. I a la inversa, un valor de rang que sigui alt i

1. Coeficient *Dice*, test *Fishers twotailed*, test exacte de *Fisher left sided*, test exacte de *Fisher right sided*, coeficient *Jaccard*, ràtio *Log-likelihood*, mesura *True mutual information*, mesura *Pointwise mutual information*, ràtio *Odds*, test khi-quadrat de Pearson, mesura *T-score*, mesura *Poisson stirling*, coeficient fi.

que vagi acompanyat d'una puntuació baixa indica que la relació que s'estableix entre les unitats que formen el candidat a terme és més habitual i, en conseqüència, és més probable que es tracti d'una combinació menys específica de l'àmbit d'especialitat o, si més no, més habitual.

clear forward signal 1 0.0389 442 903 710 4358 540 671 455
coast earth station 2 0.0363 256 279 1007 954 274 256 626
data link layer 3 0.0285 322 1589 1554 464 564 322 334
earth station antenna 4 0.028181 961 1150 279 677 85 97
earth station antennas 5 0.0263 29 961 1150 149 677 29 37
earth station Hpa 5 0.0263 29 961 1150 104 677 29 32
earth station receiver 6 0.026224 961 1150 134 677 24 35
earth station transmit 7 0.0260 21 961 1150 81 677 24 21
earth station number 7 0.0260 2 961 1150 994 677 2 42
earth station equipment 8 0.0259 51 961 1150 1648 677 57 58
earth station identification 8 0.0259 16 961 1150 291 677 16 33
earth station HPàs 8 0.0259 14 961 1150 35 677 16 16
earth station complexes 8 0.0259 12 961 1150 12 677 12 12
earth station located 9 0.0258 10 961 1150 134 677 19 10
earth station owner 9 0.0258 8 961 1150 8 677 8 8
earth station receive 10 0.0257 13 961 1150 84 677 13 13
earth station transmitter 10 0.0257 10 961 1150 42 677 10 10

FIGURA 2. Llista de candidats a terme ordenats per valor de rang

En la imatge superior (figura 2) veiem com queda reordenada la llista de candidats a terme després de ser processada per una de les tretze mesures estadístiques esmentades més amunt, concretament els resultats corresponen a la mesura *True mutual information*. Ara el primer valor que hi ha al costat del candidat a terme correspon a la informació de rang, i el següent valor correspon a la puntuació que atribueix aquesta mesura en concret al candidat en qüestió. La informació numèrica restant correspon als valors que hem obtingut en el primer pas del filtratge i que hem comentat en l'exemple anterior, és a dir, a la freqüència i al nombre de vegades que apareixen juntes en el corpus les diferents paraules que formen el candidat a terme. Observem que els candidats que tenen un mateix valor de rang també tenen una mateixa puntuació i queden ordenats consecutivament. Així, el candidat «clear forward signal» ara té un valor de rang 1, una puntuació de 0,0389 i una freqüència d'aparició en el corpus de 442.

Si revisem l'ordre en què han quedat ara reordenats els candidats a terme, veiem que candidats que quedaven situats en les primeres posicions de la llista perquè apareixien amb més freqüència en el corpus ara queden recollits en posicions més baixes que no pas candidats que abans apareixien més avall de la llista de resultats perquè tenien una freqüència més baixa; ens referim concretament als ca-

sos de «data link layer», «earth station equipment» o «earth station receive». Així, doncs, la informació de rang ajuda a precisar el caràcter més o menys terminològic que pot tenir un candidat a terme en una llista de resultats.

El filtratge dels resultats inicials fent ús de tretze mesures estadístiques ha fet possible de comparar els resultats obtinguts i alhora comprovar que hi ha unes mesures que permeten d'endreçar en ordre descendent un nombre més gran d'unitats terminològiques que no pas altres. Així, les mesures que han reordenat un major nombre de termes en les primeres posicions de la llista de resultats són el test *Fisher*, la mesura *T-score* i la mesura *True mutual information*.

3. Tècniques de recuperació d'informació aplicades a l'extracció de terminologia

En l'àmbit de la recuperació d'informació s'apliquen estratègies de localització d'unitats per identificar i classificar continguts que també són molt útils per al procés d'extracció de candidats a terme d'un corpus d'especialitat. En aquest sentit, la mesura que és força utilitzada en recuperació d'informació i que s'ha incorporat a la tasca d'extracció de terminologia és la mesura *tf-idf* (*term frequency - inverse document frequency*), que té per objectiu filtrar els termes que són presents en molts documents. En aquest plantejament, cal quantificar la freqüència d'aparició d'un terme dins un document. Aquest paràmetre, habitualment, es coneix per *factor de freqüència del terme* (*tf*, concepte local) i es considera que dóna una mesura de fins a quin punt aquest terme descriu el contingut del document, és a dir, com més vegades apareix un terme en un document, més pes semàntic té. No obstant això, els termes molt corrents gairebé no aporten la capacitat de distingir si un document és pertinent o no ho és per a una cerca concreta. Per aquest motiu, s'hi introdueix un factor calculat a partir d'una relació inversa respecte a la freqüència d'aparició del terme dins un conjunt de documents (*freqüència inversa de documents*, *idf*), és a dir, la freqüència d'aparició del terme dins un conjunt de documents decreix com més gran és el nombre de documents que en parlen; concepte basat en el corpus. I és que, com més freqüent sigui un terme en el conjunt de documents, menys pes i menys capacitat discriminatòria tindrà i, per tant, representarà, de manera secundària, el conjunt de documents. En canvi, els termes que apareixen poc en el conjunt de documents són els que tindran més pes en la mesura *tf-idf* i, per tant, representaran més bé la totalitat de documents.

En l'àmbit de l'extracció de terminologia, la mesura *tf-idf* és molt productiva per a determinar quins són els termes rellevants d'un corpus d'especialitat. Ara bé, a diferència del que es fa en l'àmbit de recuperació d'informació, la selecció de candidats a terme s'efectua fent servir un corpus de llengua general que serveix per a contrastar les unitats que apareixen en aquest corpus amb les que són

pròpies d'un corpus d'especialitat. En aquest sentit, si un candidat a terme apareix força representat i també força distribuït dins el corpus de llengua general, llavors és descartat com a possible candidat a terme. I, a la inversa, si el candidat no apareix en cap dels àmbits temàtics del corpus de llengua general, hi apareix molt poc o bé queda poc distribuït en els diferents fitxers del corpus, llavors es considera adequat com a candidat a terme. D'aquesta manera, les unitats del corpus d'especialitat que apareixen sovint i força distribuïdes en el corpus de llengua general es considera que corresponen a paraules d'ús general i no pas a paraules pròpies d'un àmbit d'especialitat i, per tant, són descartades com a unitats candidates a ser termes.

En aquest sentit, si reprenem el procés de filtratge que hem comentat més amunt tenint en compte les tècniques de recuperació d'informació aplicades a l'extracció de terminologia, el que fem ara és contrastar la llista de candidats a terme amb el contingut d'un corpus de la llengua general amb l'objectiu de poder obtenir un valor de *tf-idf* per a cada candidat.

data link layer 3.49720618070395
coast earth station 3.49720618070395
earth station number 3.49720618070395
earth station Hpa 3.49720618070395
earth station equipment 3.49720618070395
earth station transmit 3.49720618070395
earth station complexes 3.49720618070395
earth station antenna 3.49720618070395
earth station receiver 3.49720618070395
earth station antennas 3.49720618070395
earth station identification 3.49720618070395
earth station transmitter 13.49720618070395
cleara forward signal 3.49720618070395
earth station owner 3.49720618070395
earth station HPàs 3.49720618070395
earth station receive 3.49720618070395
earth station located 3.49720618070395

FIGURA 3. Llista de candidats a terme ordenats per valor de *tf-idf*

En la imatge superior (figura 3) podem observar el valor de *tf-idf* que hem obtingut per a la llista de candidats a terme amb què treballem. En aquest cas, el valor de *tf-idf* és igual per a tots els candidats, resultat que ens indica l'alt grau d'especificitat que tenen tots els candidats en ser contrastats amb un corpus de llengua general. També cal tenir en compte que són unitats que apareixen amb molta freqüència en el corpus d'especialitat; per tant, són d'aparició escassa o nul·la en un corpus de llengua general.

4. Combinació d'estratègies en el procés d'identificació d'unitats terminològiques

En el procés d'identificació d'unitats amb caràcter terminològic constatem que la combinació del valor de freqüència d'aparició d'una unitat en un corpus d'especialitat amb els valors de puntuació i rang que ens ofereixen les mesures estadístiques i el valor de *tf-idf*, que és una mesura pròpia de l'àmbit de la recuperació d'informació, permet de classificar millor la llista de candidats a terme tenint en compte el seu caràcter terminològic.

Per aquest motiu, en aquests moments avaluem la possibilitat d'establir un valor de ponderació únic que combini els quatre valors que acabem d'esmentar i, així, poder situar en les primeres posicions dels resultats les unitats que tenen un caràcter terminològic marcat i en les darreres posicions les unitats que són de caràcter menys específic. En aquest sentit, les unitats que tinguin un valor de ponderació més alt seran les que apareixeran amb molta freqüència en el corpus d'especialitat, tindran un valor de rang baix, tindran poca presència en un corpus de llengua general i se situaran en les primeres posicions de la llista de candidats a terme d'un corpus d'especialitat; aquestes unitats tindran un caràcter terminològic marcat i seran susceptibles de formar part d'una llista de termes de referència d'un corpus d'especialitat. I les unitats que tinguin un valor de ponderació més baix correspondran a unitats pròpies d'altres àmbits d'especialitat o bé a combinacions d'àmbit més general que, pel fet de ser usades en un corpus d'especialitat, poden esdevenir unitats específiques de l'àmbit. Així mateix, per poder fer una avaluació objectiva dels resultats que s'obtenen amb un valor de ponderació únic treballem amb una llista de termes de referència propis de l'àmbit d'especialitat del qual s'extreuen els candidats a terme.

En la figura 4 podem observar com queda endreçada finalment la llista de candidats a terme a partir del valor de ponderació. L'ordre en què quedaven endreçats inicialment els candidats amb la mesura *True mutual information* resulta modificat lleugerament després d'haver considerat el valor de *tf-idf* i de freqüència. A tall d'exemple, veiem que el candidat «data link layer», que segons el valor de rang de la mesura estadística *True mutual information* quedava recollit en tercera posició, ara, amb el valor de ponderació únic, queda situat en segona posició, fet que indica que té un major caràcter terminològic que no pas «coast earth station», que ara queda situat en tercera posició. O bé, «earth station Hpa», que amb el valor de ponderació queda situat més amunt en la llista de resultats que no pas amb el valor de rang o amb el valor de freqüència separadament.

clear forward signal 1
data link layer 0.8428355957776772
coast earth station 0.826395173453997
earth station antenna 0.62775263951735
earth station Hpa 0.555203619909502
earth station antennas 0.555203619909502
earth station receiver 0.518099547511312
earth station transmit 0.482503770739065
earth station equipment 0.471794871794872
earth station number 0.468174962292609
earth station identification 0.445399698340875
earth station HPàs 0.443891402714932
earth station complexes 0.442383107088989
earth station located 0.407541478129713
earth station owner 0.406033182503771
earth station receive 0.376470588235294
earth station transmitter 0.37420814479638

FIGURA 4. Llista de candidats a terme ordenats per valor de ponderació

5. CONCLUSIONS

La combinació de diverses estratègies estadístiques aplicada a l'extracció d'unitats pròpies d'un àmbit d'especialitat permet d'identificar amb més eficàcia aquest tipus d'unitats que no pas considerar els resultats obtinguts a partir d'una sola estratègia estadística. Els resultats que hem obtingut fins ara així ens ho confirmen; per aquest motiu, treballem per a poder identificar quina és la combinació de mesures estadístiques més adequada amb l'objectiu d'extreure un major nombre d'unitats terminològiques procedents de diferents corpus d'especialitat. I ho fem contrastant els resultats que ens ofereix cada mesura estadística amb els valors de freqüència, rang i *tf-idf*, tal com acabem de descriure.

En definitiva, el fet de poder identificar unitats terminològiques a partir d'un procés automatitzat facilita enormement l'elaboració de llistes de paraules clau i la construcció de taxonomies i futures ontologies en l'àmbit pròpiament de la documentació, i constitueix el material de partida per a poder plantejar un treball terminològic en el qual s'hagi de processar un gran volum de corpus en una llengua o en més d'una llengua.

6. REFERÈNCIES BIBLIOGRÀFIQUES

ARDANUY, J (2003). «Els models matemàtics de recuperació de la informació i la seva implementació en motors de cerca de propòsit general» [en línia]. A: *E-prints in Library and Information Science*. <http://eprints.rclis.org/archive/00007953/01/motors.pdf> [Consulta: 29 maig 2009].

BAEZA-YATES, R.; RIBEIRO-NETO, B (1999). *Modern information retrieval*. ACM Press.

BANERJEE, S.; PEDERSEN, T. (2003). «The Design, Implementation and Use of the Ngram Statistics Package» [en línia]. A: *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. Mèxic, p. 370-381. <http://www.d.umn.edu/~tpederse/Pubs/cicling2003-2.pdf> [Consulta: 29 maig 2009].

CHURCH, K. W.; HANKS, P (1990). «Word association norms, mutual information and lexicography» [en línia]. *Computational Linguistics*, núm. 16, p. 22-29. <http://acl.ldc.upenn.edu/J/J90/J90-1003.pdf> [Consulta: 29 maig 2009].

CODINA, L.; ROVIRA, C (2002). «Information Retrieval Techniques» [en línia]. A: *Organización y recuperación de la información*. Universitat Oberta de Catalunya. (Documents de Lectura) <http://cv.uoc.es/cdocent/BOIQM7V2N6_PVI7JZGVG.pdf> [Consulta: 29 maig 2009].