Digit·HVM revista digital d'humanitats

Sistemes de resum automàtic de documents

Salvador Climent Roca

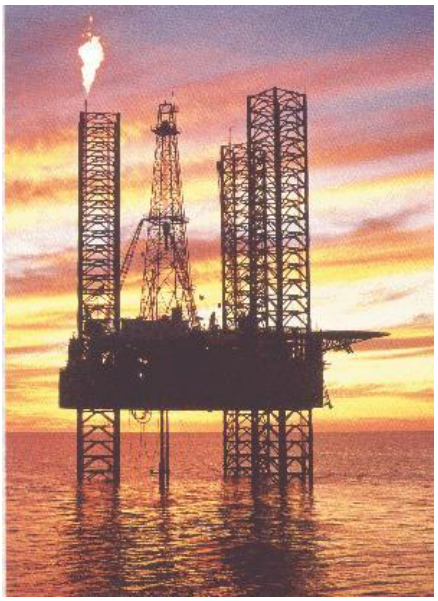
Professor dels Estudis d'Humanitats i Filologia de la UOC
scliment@campus.uoc.es

Resum: L'aparició de la WWW i d'Internet com a xarxa de telecomunicacions ha canviat el concepte del que es considera informació: no està més ben informat qui més dades té, sinó qui disposa dels millors mitjans per a obtenir exclusivament les que necessita i per a assimilar-les (digerir-les). Aquesta situació està donant un gran impuls a la recerca i el desenvolupament d'aplicacions en el camp de les tecnologies de la recuperació i l'extracció d'informació. En aquest context, els sistemes de resum automàtic de documents representen un nou pas endavant cap a una optimització del tractament de la documentació en format digital i la seva adequació a les necessitats dels usuaris. En aquest article es presenten les principals línies de recerca en la confecció de resums automàtics i les seves relacions amb altres àrees de l'enginyeria lingüística.

1. Introducció

És innecessari descriure aquí un cop més el problema que es planteja en la societat digital quant a l'obtenció, la gestió i la digestió de la informació. Tots hem llegit xifres esfereïdores sobre els milers de milions de pàgines que conté Internet, els centenars de milions de missatges de correu electrònic que s'envien anualment o els milers de milions de gigabytes d'informació nova que es genera en el mateix espai de temps (podem consultar per exemple la revista digital Infonomia).

Davant d'aquest panorama, normalment la primera reacció sol consistir en un cert desig de tirar l'ordinador per la finestra i anar-se'n a una illa deserta. Però tot i sent la més sensata, aquesta solució no acostuma a ser viable, per la qual cosa, després d'una breu reflexió, el següent pas sol ser buscar algun tipus d'ajuda: "d'entre tot aquest marasme, algú o alguna cosa pot ajudar-me a seleccionar únicament allò que m'interessa?"



Per a arribar a fer aquesta última selecció, hi ha una cosa que ens seria tremendament útil: que algú o alguna cosa ens pogués donar una idea una mica més precisa del contingut dels documents. En poques paraules, que ens respongués a la pregunta: "aquest document, de què va?".

Un ampli espectre d'investigacions en tecnologies de la informació està dirigit a la solució d'aquests problemes i d'altres similars, molts dels quals impliquen en major o menor mesura algun tipus de tractament del llenguatge humà ja que, com a ningú no se li haurà escapat, la major part de la informació que circula pel món està codificada en alguna d'aquestes curioses llengües naturals, que se'ns antullen tan fàcils i transparents quan són precisament la nostra, i tan complicades quan no ho són.

L'obtenció, el filtratge, la classificació i l'extracció d'informació, en les seves dimensions mono i multilingüe, són àmbits prioritaris de la investigació en el camp de la tecnologia lingüística. En aquest panorama, el resum automàtic de documents és possible que sigui la guinda que coronarà el pastís.

El nostre cercador favorit serà aquell al qual li puguem demanar (en la nostra llengua): busca'm documents en qualsevol idioma que parlin de tal assumpte, selecciona'm els més rellevants, classifica'ls d'acord amb tal criteri, i dóna'm un resum de cinquanta paraules de cadascun d'ells (en la meva llengua).

2. Analogies

El resum automàtic de documents comparteix o pot compartir propietats amb l'obtenció i l'extracció d'informació. De fet les dues grans línies d'investigació en aquest camp es corresponen per analogia amb aquestes dues tecnologies.

Entenem per obtenció (o recuperació) d'informació el que per exemple fan els cercadors d'Internet: d'entre un conjunt (enorme) de documents, obtenir aquells que responen a uns determinats criteris —en general, certes paraules clau—. D'altra banda, l'extracció d'informació consisteix a tractar un o diversos documents per a extreure'n una determinada informació que ens interessa i generar a partir d'ella un nou document (o una altra estructura informativa, per exemple una plantilla) que contingui únicament aquesta determinada informació rellevant. Les dues línies bàsiques d'investigació en la confecció de resums automàtics són anàlogues a aquestes.

Els resums per extracció actuen sobre un (o diversos) documents, vistos com una col·lecció d'oracions; i d'aquestes oracions extreuen i presenten aquelles considerades més rellevants o que responen a uns determinats criteris. En aquest cas el resum és un subconjunt de les oracions del text original.

Els resums per abstracció utilitzen tècniques més sofisticades de tractament del llenguatge, ja que el resultat no consisteix en determinades oracions triades del text original, sinó en un document de nova redacció generat a partir del tractament de la informació continguda el primer. Les tècniques necessàries per a l'aplicació d'aquesta estratègia encara són lluny d'haver obtingut resultats satisfactoris i pertanyen per tant al camp de la investigació bàsica. És per això que, fins al moment, els avenços més significatius s'hagin realitzat en el camp dels resums per extracció, i és a aquests a què ens referirem a partir d'aquí.

En línies generals, quan parlem d'una tecnologia tan relativament nova i prometedora, és necessari fer una prevenció per al no expert: cal no caure en l'error d'albergar falses expectatives, com va succeir en els primers temps de la traducció automàtica i en certa forma segueix succeint amb altres tecnologies com els sistemes de diàleg oral amb màquines. No hem d'esperar que sorgeixin ràpidament aplicacions perfectes, que, per així dir-ho, substitueixin els agents humans. Més aviat hem d'esperar aplicacions realistes i operatives, que ajudin els humans i els alliberin d'algunes tasques redundants i fatigoses.

Continuant amb els paral·lelismes, de la mateixa manera que la traducció automàtica no ens proporciona textos directament publicables sinó documents que han de ser invariablement posteditats, o que serveixen per fer-se una idea aproximada del seu contingut, amb els sistemes de resum automàtic normalment el que podrem obtenir seran extractes útils per saber de què va el document original, possiblement per classificar-lo en alguna estructura temàtica, i decidir si finalment ens interessa o no acudir a la lectura del text complet.

3. Tipologies

Quan hom sent parlar per primera vegada d'un programa que resumeix documents, algunes de les primeres preguntes que li vénen al cap són semblants a les següents. (1) "I aquests programes, com ho fan? De la mateixa manera que ho faríem nosaltres?". I (2) "aquests resums automàtics... són bons? Són tan bons com els que pugui fer una persona?".

Deixem per a després la resposta a la primera pregunta, però és pertinent avançar que, com és norma habitual en l'enginyeria lingüística i en general en la intel·ligència artificial, els sistemes automàtics rarament actuen seguint les mateixes estratègies que els humans. Entre altres raons perquè, en realitat, tampoc no tenim una idea molt clara de quines són les estratègies de raonament que utilitzem. Ni tampoc els sistemes informàtics disposen de cap manera del mateix tipus d'informació de què disposem les persones. En intel·ligència artificial entenem per sistema intel·ligent aquell que emula un comportament intel·ligent. O dit en altres paraules, aquell que realitza una tasca que, si l'hagués realitzat un humà, podríem dir que ha requerit d'un determinat grau d'intel·ligència.

Però comencem per la segona de les preguntes. Aquests resums automàtics, són bons? Com moltes vegades sol succeir, la primera resposta és una altra pregunta: i doncs, què és exactament un bon resum? Dependrà de diversos factors, especialment de què sigui el que hi vulguem trobar.

Potser ens interessa saber, de forma genèrica, què diu l'autor en el text; o potser el que volem és saber què diu l'autor en el text respecte a algun tema en concret. Es tracta de la diferència entre "de què parla aquest article?" i "què diu aquest article sobre la Borsa de Tòquio?".

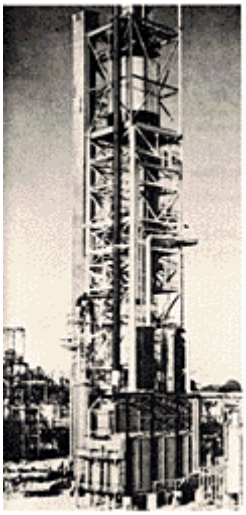
O potser la situació és que ja tenim informació prèvia sobre el tema, i volem assabentar-nos de què hi ha de nou respecte d'això. En aquest cas no volem que ens expliquin tota la història, en tenim prou amb un resum que contingui només els aspectes nous de la qüestió.

El tipus de resum que volem també pot haver de veure amb la quantitat i qualitat de la informació que desitgem obtenir. Potser ens és suficient saber més o menys de què tracta el text per poder classificar-lo, o potser el que volem és un nou document que condensi el seu contingut. En el primer cas, el cas que ens és suficient un resum purament indicatiu, potser vulguem que ens sigui ofert en forma de text coherent, o potser en tindrem prou amb una simple llista de fragments o paraules clau, o potser preferirem que el resultat sigui un indicador temàtic. En el segon cas, depenent del volum de documents i temps que tinguem per llegir-los, potser voldrem obtenir resums molt concisos, per exemple de cinquanta o cent paraules. O potser anem més folgats i podem abordar resums de fins a un parell de fulls.

I ja en una altra dimensió, potser desitjaríem que el sistema fóra prou bo com per oferir-nos un únic resum d'un conjunt de documents. Per exemple, de totes les notícies de tots els diaris d'avui que tracten de l'última crisi de govern.

Aquests exemples corresponen a algunes de les diferents tipologies i línies d'investigació actuals en aquest camp: resum indicatiu (per ser usat en classificació i filtratge de documents) vs. resum informatiu (per ser llegit en substitució o com a avanç del document complet); resum genèric vs. resum guiat (en resposta a una cerca o pregunta que defineix els interessos del lector); resum genèric vs. resum d'actualització (que ofereixi únicament els aspectes nous d'una qüestió, obviant la informació ja coneguda); resum multidocumental (que condensi en un document diversos textos que versen sobre una mateixa qüestió); resums amb diferents nivells de compressió del text original, etc. Cadascun d'aquests gèneres precisa de l'ús de tècniques diferents de tractament del llenguatge.

D'altra banda, també hem de tenir en compte que hi ha molts tipus possibles de documents a resumir: notícies, articles, columnes d'opinió, missatges de correu, reportatges, papers científics, webs, programes, llistes... potser fins i tot llibres.



I en tots els casos el nivell de marcatge del text original pot ser molt divers, des de documents en text simple sense cap tipus de marca estructural fins a documents codificats al detall en XML, passant per pàgines HTML o qualsevol altre tipus de text marcat d'una o una altra manera. El gènere i el nivell d'estructuració del text original són variables extremadament importants a tenir en compte per abordar la tasca, de manera que moltes de les tècniques utilitzades en un cas concret poden no ser en absolut reutilitzables en un altre.

Així que, en resum, què és un bon resum? Doncs bé, primer que res ens haurem de plantejar-nos el següent: un bon resum, de què, i per a què.

Però també ens havíem preguntat una altra cosa: els resums automàtics, són tan bons com els que pugui fer una persona? Doncs bé, en aquest cas és pertinent preguntar-se: "com els que pugui fer... qui?".

Una de les tasques habituals i necessàries en qualsevol tipus de recerca és l'avaluació dels resultats. En el cas del resum automàtic de textos l'avaluació del sistema se sol realitzar comparant els resums obtinguts automàticament amb uns resums tipus realitzats per humans; per a això se sol demanar a un conjunt d'analistes que realitzin resums o extractes d'un corpus de documents de prova.

I el cas és que s'ha demostrat que la pròpia obtenció d'un resum tipus no és en absolut una tasca trivial: donat un mateix document, difícilment existeix acord entre els analistes sobre quina és la informació que és pertinent seleccionar com a resum. És més, en determinades proves d'avaluació s'ha demanat a idèntics subjectes que resumeixin el contingut d'idèntics documents, havent transcorregut un cert lapse de temps entre la primera prova i la segona. Doncs bé, el grau d'acord dels avaluadors amb si mateixos va resultar ser de tan sols un 55%.

El problema se sol solucionar mitjançant mètodes estadístics, determinant d'alguna manera un resum tipus o mitjana que pugui servir de banc de proves. Sembla una manera raonable d'avaluar els sistemes de resum automàtic. No obstant això, cal que serveixin aquests comentaris simplement per a adonar-nos que una pregunta com "els resums automàtics, són tan bons com els que pugui fer una persona?", normalment no sol tenir una resposta clara.

4. Tècniques

I un cop exposats els problemes, anem a les solucions. Quines tècniques s'utilitzen per resumir automàticament?

Habitualment es classifiquen en tres famílies: tècniques basades en la superfície del text (no es realitza cap tipus d'anàlisi lingüística); tècniques basades en les entitats anomenades en el text (es realitza algun tipus de reconeixement i classificació del lèxic utilitzat); i tècniques basades en l'estructura discursiva (precisen d'algun tipus de tractament estructural del document, generalment un tractament de tipus lingüístic).

Els tractaments superficials són els que s'utilitzen habitualment en els productes comercials. Aborden el text simplement com a cadenes de caràcters —és a dir, se sol donar el cas que tracten amb lletres però també poden ser números o qualsevol altre tipus de símbol— i realitzen amb ells operacions de càlcul per seleccionar-ne alguns com a representació del document. Un mètode clàssic és la selecció dels termes estadísticament freqüents en el document. P.e., seleccionem com a resum del text les oracions (en realitat, més que d'oracions hem de parlar, més o menys, de cadenes de caràcters que comencen amb un símbol de majúscula i acaben en un punt) que contenen el major nombre de termes més freqüents en el document —tot això, amb diverses variants o precisions—.

Un altre grup de mètodes es basa en la posició. Posició en el text, en el paràgraf, en la profunditat de la classificació de seccions, etc. Es tracta de seleccionar els fragments de text que ocupen les posicions que prometen ser més rellevants. Típicament, per resumir notícies periodístiques, el millor resum curt pot ser elegir el primer paràgraf de la notícia. En articles científics, és possible que siguin especialment rellevants el compendi inicial (o *abstract*), les conclusions i la bibliografia.

Altres mètodes treuen profit de parts destacades del text: títols, subtítols, *leads*... Se suposa que les oracions que continguin les paraules del títol són millors candidates a resumir de forma adequada el document. O calculen la rellevància de les oracions recolzant-se en la presència de determinades cadenes bonificadores o penalitzadores (termes com *en conclusió*, *és important*, etc.).

Una segona família de mètodes, i entrem ja en el camp de la investigació més que no pas en el dels productes comercials, és la basada en entitats. En aquest cas ja es dóna un major nivell d'anàlisi lingüística. I dic *major* i no *algun* perquè, encara que pugui no semblar-ho, els mètodes basats a la superfície normalment sí que requereixen d'algun tipus de tecnologia lingüística: dividir un text en oracions no és una simple qüestió de majúscules i punts, perquè hi ha majúscules i punts que no delimiten oració (pensi's p.e. en *N.A.T.O.*), i cal aplicar un cert coneixement lingüístic per decidir quins caràcters (típicament, punts, però també signes d'admiració o interrogació) delimiten o no una oració.

Els mètodes basats en entitats parteixen de tècniques que permeten reconèixer unitats lingüístiques d'entre el marasme de la cadena de caràcters alfanumèrics. Quina cadena separada per espais en blanc és un nom, quina és un verb... Per fer això cal tenir analitzadors morfològics i desambiguadors lèxics, ja que una mateixa cadena pot ser nom o verb o pertànyer a una altra categoria (pensem en *casa*, com a nom o com a forma del verb *casar*). També és necessari detectar quines cadenes superficialment diferents pertanyen a un mateix concepte o categoria (pensem en*era* i *sou*, formes distintes del verb *ser*, i també per cert noms comuns); per a això és necessari disposar de programes lematizadors. I així fins a arribar a anàlisis més sofisticades de tipus sintàctic o semàntic.

Aquest tipus de mètodes poden ser capaços de prendre les entitats (prèviament analitzades lingüísticament) i detectar diversos tipus de relacions establertes entre elles: recurrència de formes o lemes, relacions semàntiques (un cotxe, una moto i un autobús són vehicles), relacions temàtiques (un àrbitre, un davanter centre i un penal són termes de l'àmbit del futbol), etc.

A partir d'aquí és possible construir una representació de la connectivitat del text, en termes de grafs, de manera que sigui possible determinar quines parts del text (oracions) són especialment rellevants i per tant són candidates a formar part de l'extracte o resum.

Per implementar aquest tipus de mètodes és especialment important comptar amb bases de coneixement lèxic (com EuroWordNet), reconeixedors d'entitats (p.e. de noms propis, paraules que per la seva pròpia naturalesa normalment no formen part dels diccionaris o repertoris lèxics) i sistemes de resolució de referències anafòriques (p.e. d'anàfores pronominals).

Finalment, els mètodes basats en l'estructura poden treure partit de la pròpia carcassa hipertextual d'un document HTML, o en un altre àmbit de coses poden intentar, partint de la base de les tècniques i recursos d'enginyeria lingüística abans apuntats i altres de tipus més sofisticat, construir una estructura de l'argumentació continguda en el document que permeti detectar els fragments discursivament més rellevants. Aquest últim tipus de tècniques es basa en la detecció de marcadors discursius, bàsicament conjuncions i adverbis, del tipus *en primer lloc*, *al contrari*, *no obstant això*, *a més a més*, etc. Cal insistir

en què aquest tipus de mètodes pertany encara al més estricte camp de la investigació.

Un aspecte important a tenir en compte per a tots els mètodes basats en extracció és el de la recomposició del text. És evident que quan hom crea un document tot destriant oracions del text original és molt fàcil que es produeixin disfuncions en les referències i, en general, en la cohesió textual. Un bon sistema basat en extracció haurà de ser capaç de, a partir de la comparació i l'anàlisi del text triat i el text original, inserir material lèxic que recompongui les cadenes de cohesió textual.

5. Referències i productes

Si volem conèixer l'estat actual de la qüestió al camp del resum de documents haurem d'atendre d'una banda als sistemes comercials existents, i d'una altra a l'estat de la recerca.

En el primer cas podem acudir al producte d'Inxight (Xerox), al més recent deCopernic, o al típic i elemental sistema d'autoresum de Microsoft Word. Un llistat prou complet de sistemes actualment accessibles el podem trobar en la pàgina deMitre Org. Quant a sistemes en vies de desenvolupament que operen online, podem posar a prova els molt interessants SweSum i Extractor (de fet, la tecnologia que hi ha darrera el resumidor de Copernic), que ofereixen resums de l'espanyol.

Quant a l'estat de la investigació, són molt rellevants les pàgines de la Universitat d'Ottawa, de l'investigador de la Universitat de Columbia Dragomir Radev i, sobretot,l'extensa bibliografia continguda en la pàgina que manté el propi Radev a la Universitat de Michigan. A l'Estat espanyol, estem construint un prototip de sistema resumidor de notícies periodístiques en el marc del projecte Hermes.

Però els programes de major envergadura cal localitzar-los en el marc de les accions impulsades pel Departament de Defensa dels EE.UU., DARPA, com el programaTIDES per a la detecció, extracció i resum d'informació multilingüe, i el congrés de comprensió de documents, Document Understanting Conference, DUC, que proposa una sèrie d'experiments a gran escala per a l'avaluació de tècniques i sistemes de resum i comprensió de text.

En tot cas, i per finalitzar, és important remarcar novament que per superar els sistemes comercials ja existents i avançar cap a sistemes relativament intel·ligents o posseïdors d'un cert coneixement, és crucial comptar amb una gran bateria de recursos lingüístics, com analitzadors morfològics, desambiguadors categorials, bases de coneixement lèxic, analitzadors sintàctics, reconeixedors d'entitats o sistemes de resolució d'anàfores i altres tipus de coreferencia. Com sol succeir en la majoria de tecnologies de la llengua, el creixement qualitatiu és de tipus asintòtic, de manera que és relativament fàcil aconseguir nivells de funcionament acceptables però, de forma progressiva, cada cop és més costós elevar els nivells de qualitat dels sistemes.

A aquest respecte, és imprescindible el treball en equips interdisciplinaris, en els que lingüistes i enginyers informàtics col·laborin estretament, ja que en el terreny de la tecnologia lingüística, a diferència de moltes altres àrees d'investigació en lingüística i altres ciències humanes i socials, els esforços individuals, per meritoris que resultin, rarament condueixen a resultats acceptables.

6. Apèndix

6.1. Resum de l'article, de 100 paraules, realitzat per SweSum (en castellà)

Aquest resum ha estat generat automàticament pel sistema SweSum, i es presenta aquí sense cap retoc. S'ha presentat per extracció a partir del text original d'aquest article en espanyol, ja que no es disposa actualment de cap sistema de resum operatiu pel català.

> Para exponerlo con ejemplos paralelos, del mismo modo que la traducción automática no nos proporciona textos directamente publicables sino documentos que deben ser

invariablemente posteditados, o que sirven para hacerse una idea aproximada de su contenido, con los sistemas de resumen automático normalmente lo que podremos obtener serán extractos útiles para saber de qué va el documento original, posiblemente para clasificarlo en alguna estructura temática, y decidir si finalmente nos interesa o no acudir a la lectura del texto completo. Y (2) "esos resúmenes automáticos... Esos resúmenes automáticos, ¿son buenos?

6.2. Resum de l'article, de 200 paraules, realitzat per Extractor (en anglès)

Aquest resum ha estat generat automàticament pel sistema Extractor, i es presenta aquí sense cap retoc. S'ha realitzat per extracció a partir del text original d'aquest article en anglès, ja que no es disposa actualment de cap sistema de resum operatiu pel català.

- **Retrieving**, managing and digesting information are hard and crucial tasks in our society; we therefore call for systems and applications able to help us in such heavy duties: given the billions of pages the Web comprises, please help me to select just that precise information which I am interested in.

- A wide range of research on **language technologies** focuses on such and related problems, many of them involving to a greater or smaller extent some kind of human language processing inasmuch as most of the information flowing around the world, as anyone can notice, happens to be codified in a natural language.

- Information Retrieval, Filtering, Classification and **Extraction**, either in their monolingual or multilingual dimensions, are key fields of research in linguistic technologies.

- In this environment**, automatic summarization** of documents may probably be the finishing touch.

- Summaries by abstraction, i.e. abstracts, make use of rather more **complex linguistic technology techniques**, as long as the output is not a simple number of sentences retrieved as found in the input text but instead a brand new document which has been generated by processing the information contained there.

- One should not expect ready-to-use perfect systems able to replace **humans** in their work.

Enllaços relacionats:

- DARPA Information Technology Office: Translingual Information, Detection, Extraction and Summarization (TIDES): http://www.darpa.mil/ito/research/tides/
- The Text Summarization Project. Universitat d'Ottawa: http://www.site.uottawa.ca/tanka/ts.html
- A Bibliography of Research in Text Summarization: http://www.si.umich.edu/~radev/summarization/large-bib.doc
- SweSum: http://www.nada.kth.se/~xmartin/swesum/index-eng.html
- Extractor: http://extractor.iit.nrc.ca/on_line_demo.html

[Data de publicació: setembre de 2001]

Digit·HVM revista digital d'humanitats

Citació recomanada:

CLIMENT ROCA, Salvador (2001). “Sistemes de resum automàtic de documents”. *Digithum*, núm. 3 [article en línia]. DOI: http://dx.doi.org/10.7238/d.v0i3.598