RESUMIDOR DE NOTICIES EN CATALÀ DEL PROJECTE HERMES

Maria Fuentes, Edgar Gonzàlez i Horacio Rodríguez
Centre de recerca TALP
Universitat Politècnica de Catalunya
{mfuentes, egonzalez, horacio}@lsi.upc.es

RESUM
En aquest document es presenta un sistema que permet resumir automàticament documents escrits en català, via extracció de fragments. Aquest sistema explota les propietats de cohesió del text a través de la detecció de cadenes lèxiques, de coreferència i d'entitats nominals. El sistema s'ha utilitzat per resumir noticies d'agència en català, castellà i anglès. En aquest article es presenta l'avaluació del sistema per al català, comprovant que el funcionament del sistema és independent de si els documents a resumir són escrits en català o en castellà.

INTRODUCCIÓ

L'objectiu d'aquest estudi és l'avaluació del sistema de resum automàtic de notícies en català desenvolupat dins el projecte HERMES[1]; en el marc del qual s'han construït diverses aplicacions d'accés a informació textual multilingüe (per al català, castellà, euskera i anglès).
Basant-se en les propietats de cohesió del text, el sistema extreu els fragments més informatius del document. Per reflectir la cohesió existent entre les paraules que componen un text utilitzem el concepte de cadena lèxica, és a dir, no considerem que el text és un conjunt de paraules independents entre si, sinó que entre elles existeix una relació a nivell lèxic. Amb l'objectiu de crear un sistema independent de la llengua, la similitud entre paraules es determina mitjançant l'ús d'EuroWordNet[2], base de dades lèxica multilingüe (català, castellà, euskera, anglès, francès, italià, holandès, txec, estonià i alemany). Aquestes xarxes de paraules s'estructuren en conjunts de sinònims, entre què s'estableixen una sèrie de relacions semàntiques bàsiques. De manera complementària utilitzem cadenes d'entitats nominals i coreferents. La idea és que cada cadena sigui indicadora d'una línia temàtica.
Els experiments presentats fan referència a l'aplicació del sistema en el resum de notícies de l'agència EFE. Degut al cost que suposa la creació d'un corpus de test, per avaluar el funcionament del sistema en català s'han traduït automàticament del castellà al català els documents i resums de test creats per avaluar el sistema en castellà.

Aquest document s'estructura de la següent manera: un cop feta aquesta introducció, es defineix el concepte de cadena lèxica, i a continuació es presenta l'arquitectura del sistema. Tot seguit els experiments realitzats i els resultats obtinguts són detallats i per acabar s'exposen les conclusions a les què s'ha arribat amb aquest treball.

---

[1] Hemerotecas Electrónicas, Recuperación Multilingüe y Extracción Semántica (http://terral.ieec.uned.es/hermes)
[2] Pels experiments en català i castellà hem utilitzat EuroWordNet (http://www.hum.uva.nl/~ewn/WordNet). Per l'anglès també pot ser utilitzada la WordNet de Priceton (http://www.cogsci.princeton.edu/~wn/w3wn.html

CADENES LÈXIQUES

Les cadenes lèxiques permeten relacionar diferents oracions o parts de text. Aquest terme va ser introduït per [6]. A [7] es diferencien dos tipus de cadenes: les d'identitat i les de similitud. Les primeres contenen termes que fan referència a un mateix objecte i són creades per cohesió pronominal, repetició lèxica o instàncies equivalents. Aquestes cadenes estan estretament lligades al contingut del text, ja que les relacions de coreferència poden ser determinades només pel context. En canvi, entre els elements de les cadenes de similitud, relacionats semànticament, es produeix un lligam supra-textual. La Figura 1 mostra un exemple de noticia d'agència EFE original en català i algunes de les seves cadenes lèxiques.

Ordino (**Andorra**), 7 jun (EFE).- L'Auditori Nacional d'Ordino ha acollit aquesta tarda la presentació de l'obra audiovisual seriada en dos capítols de 90 minuts i titulada "**Andorra**, entre el Torb ***i*** la Gestapo", basada en la novel·la del mateix títol de Francesc Viadiu Vendrell, produïda per Ovideo i finançada pel Govern andorrà i Televisió de Catalunya.

L'acte ha comptat amb la presència, entre altres personalitats, del cap de govern d'**Andorra**, Marc Forné, i del ministre de Cultura, Enric Pujal, així com del conseller de Cultura de la Generalitat, Jordi Vilajoana, i el director de TV3, Lluís Oliva.

Així mateix ha assistit el director de la **producció**, Lluis Maria Güell, i els dos actors protagonistes, ***Antoni Valero*** i ***Mónica López***, a més d'altres actors que intervenen al **film**.

L'acció de la **sèrie**, basada en fets reals, es desenvolupa a **Andorra** durant la II Guerra Mundial, quan el petit país pirinenc es va convertir en centre d'una xarxa de passada d'aviadors aliats caiguts en territori francès controlat pels alemanys.

Lluis María Güell és director-realitzador des de 1971 i la seva trajectòria comprèn programes dramàtics, **pel·lícules** de televisió, minisèries ***i*** musicals, entre altres, tant en suport videogràfic com cinematogràfic. També ha realitzat vídeos institucionals i publicitaris, i ha dirigit espectacles culturals.

De la seva banda, ***Antoni Valero*** té una àmplia trajectòria en el món del teatre, el cine i la televisió, destacant el seu paper en la popular **sèrie** "Mèdic de Família", mentre que ***Mónica López*** ha participat en obres de teatre, **pel·lícules** i **sèries** televisives tan populars com "Oh, Europa" ***i*** "Nissaga de poder".

cadena lèxica forta o significativa (ex. Cadena similitud) **pel·lícules**

cadena d'entitat nominal o significativa (ex. Cadena identitat) **Andorra**

Figura 1. Exemple de noticia d'EFE original amb alguna de les seves cadenes lèxiques.

ARQUITECTURA DEL SISTEMA

L'esquema de l'arquitectura del sistema, presentat a la Figura 2, es pot dividir en tres blocs: el que determina les característiques lingüístiques, el que calcula les cadenes lèxiques, d'Entitats Nominals i les referèncials i, per últim, l'encarregat de puntuar i seleccionar els fragments que formen part del resum. Aquest sistema és pràcticament idèntic al presentat i avaluat per al castellà a [4]. La diferència bàsica, en aquest treball, és l'ús de la versió catalana del classificador d'Entititats Nominals i de l'analitzador morfològic actualitzat [3]. En general, les versions d'aquestes eines segueixen models diferents per a cada llengua, i això fa que els etiquetats no sigui exactament igual en castellà que en català. Per obtenir la informació semàntica, necessària per computar les cadenes, es fa servir la versió catalana de WordNet. Un cop preprocessats els documents, s'identifiquen i ponderen les cadenes, destacant-ne les *cadenes fortes*, és a dir, les considerades més significatives. A diferència del treball què es basa la nostra proposta [1], a part dels noms comuns, el nostre sistema permet tenir noms propis, entitats nominals, sintagmes nominals definits i pronoms com a possibles integrants d'una cadena. Per altra banda el nostre sistema permet tractar documents en català i en

castellà, a part d'anglès, essent extensible a qualsevol altre llengua, sempre i quan disposem dels recursos requerits en el preprocés: com a mínim un analitzador morfològic i la versió de WordNet per a aquesta llengua.

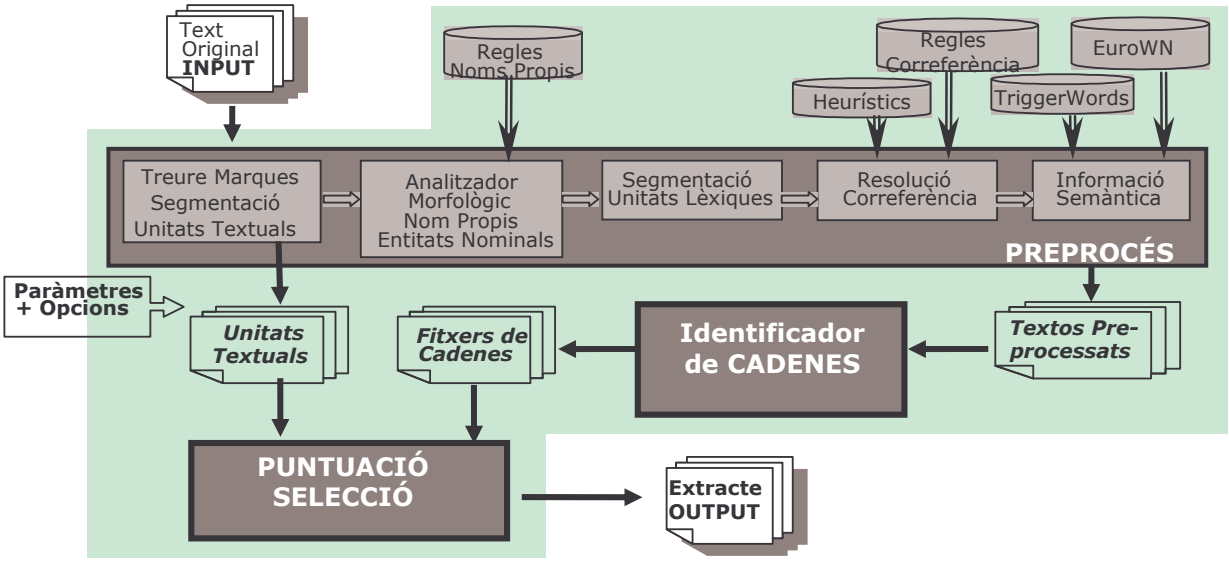

Figura 2. Arquitectura del sistema

Finalment al tercer pas, es puntuen i seleccionen els segments que formaran part del resum. A part de la llengua, el sistema té un conjunt de paràmetres, permetent definir característiques com la unitat de fragmentació (sintagma, oració, paràgraf), el grau de compressió a aplicar (%, núm. de paraules) o el tipus de resum (un sol o diversos documents). Per altra banda també dóna l'opció de calcular les cadenes i de ponderar els fragments de diverses maneres.
Actualment el sistema s'ha integrat amb altres eines de processament lingüístic, seguint la proposta [5], dins el marc del projecte CHIL[3]. Aquest projecte té com a objectiu el processament de text generat per un sistema de reconeixement automàtic de veu. Així doncs podem dir que el sistema que presentem ha estat adaptat permetent resumir tant veu com text.

AVALUACIÓ DEL SISTEMA

Amb l'objectiu de comprovar que el sistema és independent de la llengua, s'ha reproduït l'experiment fet per avaluar el sistema en castellà. S'ha utilitzat el mateix corpus d'avaluació creat per al castellà: 120 noticies de temàtica diversa, reduït a 111 un cop eliminades les noticies d'un sol paràgraf. 31 persones participaren en la creació dels resums. L'objectiu era l'obtenció d'un mínim de 5 resums per article, produïts per humans diferents.

	ORIGINAL	TRADUCCIÓ
Noms Propis	Francisco Arias Milla Cuauhtémoc Cardenas Carlos Ruiz Sacristán	Francisco **Àries** Milla Cuauhtémoc **Moradenques** Carlos Ruiz **Sagristà**
Entitats Nominals (equips beisbol)	Trotamundos de Carabobo Los Padres de San Diego	**Rodamón** de Carabobo Els **Pares** de **St Dídac**
Acrònims	(EFE) (PAN)	**(EFA)** **(PA)**

Taula 1. Exemples de traduccions incorrectes

[3] Computers In the Human Interaction Loop (http://chil.server.de/)

Degut al cost que suposa la creació d'un corpus d'aquest estil, l'avaluació per al català s'ha realitzat utilitzant el sistema de traducció automàtic castellà-català interNOSTRUM [2], per traduir prèviament el corpus. S'ha considerat que el funcionament del sistema és suficientment bo per a les nostres necessitats. De fet, analitzant les traduccions produïdes l'únic problema que detectem és en algunes de les traduccions de noms propis, entitats nominals i acrònims (veure Taula 1). De totes maneres pensem que aquest problema no té perquè afectar massa el funcionament del sistema.
Com per al castellà, degut a les característiques dels resums utilitzats en el test, la unitat d'extracció que s'ha utilitzat és el paràgraf, tot i que s'han realitzat altres experiments, no detallats en aquest document, amb segments de text més petits. També s'han provat diferents grau de compressió, però en aquest document només parlem dels corresponents al corpus de test.

MESURES	CATALÀ	CASTELLÀ
Precisió	0.83	0.85
Cobertura	0.83	0.85
Cosinus Simple	0.86	0.88

Taula 2. Resultats dels experiments

La Taula 2 mostra els resultats de dues execucions del resumidor, una per als documents en català i l'altra per als documents en castellà. Per avaluar el nostre sistema, comparant-lo amb el corpus de resums creats manualment, hem utilitzat les mesures de precisió i cobertura, així com la del cosinus simple, del paquet d'avaluació MEADeval[4], desenvolupat en el projecte MEAD. Degut a les característiques dels documents del corpus (noticies d'agència) s'ha vist que el millor resum s'obté seleccionant el primer paràgraf del document; per aquesta raó la mesura del cosinus simple és la més significativa, ja que té en compte les paraules que conté el resum i no només el paràgraf seleccionat.
Ni en català ni en castellà s'han tingut en compte cadenes referencials, és a dir, només s'han identificat cadenes formades per noms comuns, noms propis i entitats nominals. Per altra banda, les cadenes significatives s'han calculat considerant les relacions semàntiques extra-fortes (repeticions) i fortes (sentits de wordnet relacionats a distància 1). I per últim, l'heurístic aplicat per seleccionar el segment més rellevant ha estat el de donar prioritat al primer fragment creuat per alguna cadena significativa.

```
1_59 Francesc_Viadiu_Vendrell Francesc_Viadiu_Vendrell NP00SP0 PERSON * *
1_62 , , Fc * * *
1_63 produïda produït AQ0FSP * * *
1_64 per per SPS00 * * *
1_65 Ovideo Ovideo NP00G00 LOCATION * *
1_66 i i NCFS000 * * *
1_67 finançada finançat AQ0FSP * * *
1_68 pel pel SPCMS * * *
1_69 Govern Govern NP00O00 ORGANIZATION * *
```

Figura 3. Part del text de la Figura 1 preprocessat, amb un parell d'errors

Un cop estudiades detingudament les diferències entre les execucions per al castellà i català, veiem que en la majoria dels casos els errors es produeixen per no arribar a ser reconegudes exactament les mateixes cadenes significatives en català que en castellà. A la Figura 1 podem

[4] http://perun.si.umich.edu/clair/meadeval

veure que s'ha marcat *"i"* com a cadena, això és exemple d'un mal etiquetatge morfològic (veure Figura 3). En aquest cas s'etiqueta *i* com a Nom Comú.
Per altra banda, un altre tipus de problema que es dóna és en la classificació d'entitats nominals. La Figura 4 mostra les diferents classificacions que es donen per les diverses aparicions d'una mateixa paraula en un document.

```
2_58 Exèrcit Exèrcit NP00SP0 PERSON * *
3_91 Exèrcit Exèrcit NP00000 OTHERS * *
6_262 Exèrcit Exèrcit NP00G00 LOCATION * *
7_301 Exèrcit Exèrcit NP00000 OTHERS * *
7_342 Exèrcit Exèrcit NP00O00 ORGANIZATION * *
```

Figura 4. Diverses aparicions d'una entitat nominal en un document i les classificacions.

Podem considerar que les diferències entre llengües majoritàriament són degudes a una propagació dels errors introduïts en la primera etapa, la del preprocés lingüístic. De totes maneres cal dir que aquest tipus d'errors tant es produeixen per al català com per al castellà.

CONCLUSIONS

Hem presentat un sistema de resum automàtic independent de la llengua, basat en la cohesió del text, i n'hem avaluat el funcionament amb textos de noticies d'agència en català. Per fer aquesta avaluació hem utilitzat un corpus traduït automàticament de manera satisfactòria per a la nostra tasca i hem vist que la qualitat del resultat pot veure's afectada per la qualitat de les eines del preprocés lingüístic.
Actualment el sistema s'està estenent en diverses direccions i vol ser utilitzat per resumir veu.

BIBLIOGRAFIA

[1] R. Barzilay. 1997. Lexical Chains for Summarization. Master thesis. Ben-Gurion University of the Negev.

[2] R. Canals-Marote, A. Esteve-Guillen, A. Garrido-Alenda, M. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P. Perez-Antón, M. Forcada. 2001. The spanish-catalan machine translation system internostrum. Publicat a MT Summit VIII: Machine Translation in the information Age, Santiago de Compostela, Espanya.

[3] X. Carreras, L. Chao, L. Padró, M. Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. Publicat a LREC-2004, Lisboa, Portugal.

[4] M. Fuentes, H. Rodríguez. 2002. Using cohesive properties of text for automatic summarization. Publicat a JOTRI-2002, València, Espanya.

[5] E. Gonzàlez i Pellicer. 2004. Un sistema genèric de cerca de la resposta, Projecte Final de Carrera. Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya, 2004

[6] M. A. K. Halliday, R. Hasan. 1976. Cohesion in English. English Language Series. Longman Group Ltd, 1976.

[7] R. Hasan. 1984. Understanding Reading Comprehension, chapter Coherence and Cohesive Harmony. IRA: Newark, Delaware 1984.

AGRAÏMENTS

Aquesta recerca s'ha realitzada amb recursos dels projectes ALIADO (TIC2002-04447), del departament espanyol de recerca, i CHIL (IST-2004506969). El TALP és reconegut com un Grup de Recerca de Qualitat (2001 SGR 00254) pel DURSI.