UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Resum abstractiu de noticies basat en xarxes neuronals

TREBALL FI DE GRAU

Grau en Enginyeria Informàtica

Autor: Pere Marco Garcia

Tutor: Encarnación Segarra Soriano
Lluís Felip Hurtado Oliver

Curs 2020-2021

Resum

La quantitat d'informació que disposem actualment no ha fet més que incrementar enormement a les últimes dècades. En conseqüència, disposem d'un gran nombre de textos dels quals cal determinar la utilitat per a qualsevol activitat concreta. En aquest context, l'existència de resums adjunts a aquests textos, com pot ser el cas dels articles d'investigació o les notícies d'alguns periòdics, resulten de gran utilitat per agilitzar aquesta tasca. D'acord amb aquest últim fet extraiem el benefici que suposaria tindre accés a resums fidedignes dels textos que trobem a internet.

En el present treball es tractarà d'elaborar resums d'un conjunt de noticies tant en català com en espanyol. En aquest procés, s'utilitzaran diferents tecnologies per a la creació dels resums mitjançant tècniques tradicionals no supervisades i sistemes basats en xarxes neuronals. Aquests resums s'avaluaran mitjançant mètriques de caràcter sintàctic i semàntic per poder comparar la qualitat dels resums.

Finalment, s'obtindran els resums extrets mitjançant l'ús de xarxes neuronals, s'analitzaran els resums generats i es contrastaran amb els textos de referència.

Paraules clau: resum de textos periodístics, DACSA, word embeddings, resum abstractiu, resum extractiu

Resumen

La cantidad de información de la que disponemos actualmente no ha hecho más que incrementar enormemente en las últimas décadas. En consecuencia, disponemos de un gran número de textos se los que es necesario determinar su utilidad para cualquier tarea concreta. En este contexto, la existencia de resúmenes adjuntos a estos textos, como puede ser el caso de artículos de investigación o las noticias de algunos periódicos, resultan de gran utilidad para agilizar dicha tarea. En base a este último hecho, extraemos el beneficio que supondría tener acceso a resúmenes fiables de los textos que encontramos en internet.

En el presente trabajo se tratará de elaborar resúmenes de un conjunto de noticias tanto en catalán como en español. En este proceso se utilizarán diferentes tecnologías para la creación de resúmenes mediante técnicas tradicionales no supervisadas i sistemas basados en redes neuronales. Estos resúmenes se evaluarán mediante métricas de carácter sintáctico i semántico para poder comparar la calidad de los resumenes.

Finalmente, se obtendrán los resúmenes extraídos mediante el uso de redes neuronales, se analizarán los resúmenes generados y se contrastarán con los textos de referencia.

Palabras clave: resumen de textos periodísticos, DACSA, word embeddings, resumen abstractivo, resumen extractivo

Abstract

The amount of information available to us today has only increased enormously in recent decades. Consequently, we have a large number of texts which it is necessary to determine their usefulness for any specific task. In this context, the existence of summaries attached to these texts, such as research articles or news from some newspapers, are very useful to speed up this labour. Based on this last fact, we extract the benefit of having access to reliable summaries of the texts that we find on the internet.

In this work we will try to prepare summaries of a set of news in both Catalan and Spanish. In this process, different technologies will be used to create summaries using traditional unsupervised techniques and neural network based systems. These results will be evaluated using syntactic and semantic metrics to be able to compare the quality of the summaries.

Finally, the abstracts extracted through the use of neural networks will be obtained, the results generated will be analyzed and they will be contrasted with the reference texts.

Key words: journalistic text summarization, DACSA, word embeddings, abstractive summarization, extractive summarization

Índex

Índex de figures

Índex de taules

CAPÍTOL 1
Introducció

Amb la utilització cada vegada més popular d'Internet i els seus recursos la quantitat d'informació de que disposem, tant a escala d'usuari com en els àmbits professionals no para de créixer de manera exponencial. En l'actualitat és necessari en la majoria d'àmbits confiar en el capacitat dels buscadors de proporcionar la informació que busquem. Tot i això, la quantitat de recursos que se'ns presenta continua sent inabastable per a poder analitzar en un temps raonable si tota la informació és d'utilitat.

En aquestes situacions, el resum automàtic de textos suposa un gran avantatge per alleugerir la carrega que suposa la tasca d'analitzar tota aquesta informació. El propòsit final de la generació automàtica de resums és tornar un text considerablement més breu que l'original però mantenint les idees i aspectes principals. En conseqüència resulta més senzill determinar si un text pot servir per a una determinada feina. Açò pot resultar d'utilitat en una gran varietat d'àmbits en què es pot trobar un gran volum d'arxius de text com poden ser noticiaris de periòdics, transcripcions de retransmissions o bibliografia mèdica [12].

En aquest treball ens centrarem en el tractament d'un conjunt de notícies en català i en castellà. El conjunt de notícies que es treballarà forma part del corpus DACSA. Es tracta d'un corpus de notícies extretes de pàgines web periodístiques que tenen associades resums proporcionats pels mateixos mitjans de comunicació que ens serviran com resums de referència durant l'experimentació i avaluació. DACSA ha sigut desenvolupat per un grup d'investigació, l'ELiRF, amb el que s'ha treballat durant el transcurs d'una beca col·laboració atorgada pel Ministeri d'Educació i Formació Professional. Aquest corpus s'estudiarà i serà utilitzat per a la generació de resums mitjançant l'aplicació de diferents tècniques de resum automàtic de textos, tant amb tècniques que fan ús de xarxes neuronals com mitjançant mètodes tradicionals. Finalment observarem el comportament dels resums generats mitjançant l'ús de diferents mètriques i s'avaluaran els resums.

1.1 Motivació

Des que l'ésser humà va començar a ser racional s'hi ha qüestionat la naturalesa de la intel·ligència i quin és el funcionament d'aquesta i com un conjunt de material inert és capaç d'interactuar i aconseguir percebre, entendre i fins i tot predir el comportament i la informació que proporciona un món més complex que ell mateix. El camp de la intel·ligència artificial intenta no només comprendre aquest mecanisme sinó recrear-lo [36].

D'igual manera que la paraula intel·ligència pot donar pas a moltes i molt diverses accepcions, existeix un debat entre allò que realment defineix el concepte d'intel·ligència artificial. Tot i això, les principals postures defenen que es tracta de l'emulació o re-

creació d'una actuació racional o d'acord amb el que caldria esperar del comportament humà. Algunes de les disciplines més importants que es desenvolupen a partir de la intel·ligència artificial són la recreació de sistemes per a la realització de tasques concretes, la simulació de sentiments, la cognició humana i el calc de la capacitat creativa.

Entre aquestes tres tasques l'aprenentatge resulta un element comú important. El procés d'aprenentatge suposa de major rellevància a l'àmbit de la informàtica, ja que, a diferència dels éssers humans, els ordinadors no disposen dels mateixos elements que faciliten l'adquisició de coneixements d'una manera implícita. L'aprenentatge és d'una activitat que implica la presència, en major o menor mesura, de distints fets o accions. Principalment es destaquen l'existència d'un coneixement declaratiu de la matèria, el desenvolupament de capacitats cognitives que permeten assimilar el coneixement mitjançant el seguiment de metodologies o procediments, l'habilitat d'obtenir dades concretes a partir d'informació general, la utilització d'una representació efectiva i el plantejament i comprovació de noves hipòtesis a través de l'experimentació i observació. Tots aquests aspectes s'han traslladat a la informàtica i són clarament identificables en el que coneixem actualment per aprenentatge automàtic [28].

L'aprenentatge automàtic és una disciplina de la informàtica que s'ha desenvolupat des de la segona meitat del segle passat i ha guanyat rellevància en pràcticament tots els àmbits de la ciència i ha tingut impacte en la societat. La diversitat de tasques per a les quals es pot utilitzar va des de l'elaboració de diagnòstics mèdics, l'anàlisi de cadenes d'ADN i l'optimització de processos industrials fins a la predicció de comportaments humans, el reconeixement de la parla i l'anàlisi de textos [24].

Aquestes dues últimes disciplines formen part de l'àrea del Processament del Llenguatge Natural (PLN). Aquesta consisteix en un ample ventall de tècniques computacionals dirigides a l'anàlisi i representació automàtiques del llenguatge humà. Des del seu naixement a la dècada dels anys cinquanta, la investigació del PLN s'ha centrat en tasques com la cerca de respostes, la traducció automàtica, la recuperació d'informació, la classificació de textos, la generació automàtica de resums de documents i, de manera més recent, l'anàlisi de sentiments. Aquesta investigació al llarg dels anys s'ha centrat principalment en l'àmbit sintàctic, ja que era una via més directa d'abordar la majoria de problemes. Un enfocament semàntic en la resolució dels problemes del PLN sembla prometre millors resultats en la majoria de disciplines en les quals es treballa. Tot i això, fins fa relativament poc aquesta via ha quedat en un segon pla principalment per la complexitat que suposa aquest tipus de solucions i per la dependència que té dels grans volums de dades que es dispose [7].

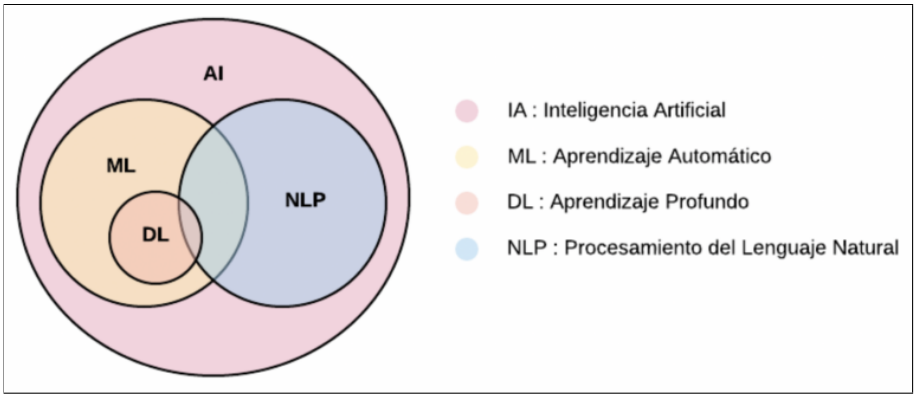

Figura 1.1: Classificació del PLN en la Intel·ligència artificial: l'estudi del PLN es superposa amb altres camps d'estudi dins de la mateixa intel·ligència artificial com el *Machine Learning* i el *Deep Learning*.

Actualment, mitjançant l'ús de les xarxes neuronals junt amb el desenvolupament del *Machine Learning*[1] i el *Deep Learning*[2] l'aproximació semàntica d'aquesta àrea de la Intel·ligència artificial està guanyant cada vegada més popularitat i aconseguint millors resultats en certes disciplines que la via basada en la dimensió sintàctica. En l'actualitat s'han desenvolupat xarxes especialitzades a treballar els textos enfocats en la semàntica conjuntament amb estructures de dades que s'adapten de manera apropiada a les necessitats dels problemes.

La creació, execució i anàlisi d'eines destinades al PLN es troba en un moment de creixement a causa de les innovacions que es publiquen arreu del món. La generació automàtica de textos no n'és l'excepció, ja que en els últims anys s'han desenvolupat sistemes, com són BART o PEGASUS, amb utilitats en aquest camp que ofereixen resultats prometedors per a aquesta nova tendència. A més a més, concretament l'obtenció de resums a partir de conjunts de notícies resulta una àrea altament explorada, sobretot en parla anglesa. En conseqüència en els últims anys es disposa d'un gran recull de referències en les quals recolzar-se per a realitzar una experimentació adequada i amb la possibilitat de comparar-la i analitzar-la correctament. Per altra banda, la possibilitat d'utilitzar ferramentes d'aquest tipus en textos en idiomes menys explorats com són l'espanyol i, sobretot, el català suposa una oportunitat de tindre resultats rellevants que puguen suposar una aportació transcendent per a aquest camp de recerca.

1.2 Objectius

El present treball té l'objectiu d'utilitzar diferents tècniques de generació de resums i analitzar i comparar els resums resultants sobre un corpus de noticies en català i castellà. Per a la realització de la labor podem distingir els següents subobjectius:

1. **Anàlisi de les característiques pròpies del corpus.** Serà necessari el coneixement d'algunes propietats del conjunt de dades per a la correcta utilització de les tècniques de generació de resums. També és d'utilitat aquesta informació per tal de poder conéixer les seues característiques.

2. **Generació de resums fent ús de tècniques tradicionals no supervisades.** Donat que no hi ha treballs relacionats que treballen sobre aquest corpus és necessari generar resums mitjançant distintes tècniques no supervisades amb la finalitat d'establir unes cotes de partida (*baseline*).

3. **Utilitzar sistemes basats en xarxes neuronals per a la generació de resums.** Es pretén utilitzar diferents tècniques basades en xarxes neuronals sobre el conjunt de dades per tal d'obtenir resums de caràcter extractiu i abstractiu.

4. **Avaluar els resums generats en els distints processos.** Mitjançant l'ús de diferents mètriques s'obtindran els valors associats als resums generats pels diferents sistemes. Amb aquests valors podrem conéixer propietats relacionades, entre altres, amb la similitud sintàctica o semàntica dels resums.

[1] Disciplina de la Intel·ligència Artificial que s'encarrega d'explorar i modelar mètodes d'aprenentatge de manera automàtica [28].

[2] Especialització del *Machine Learning* que fa ús de tècniques avançades relacionades, sobretot, amb les xarxes neuronals.

1.3 Objectius de desenvolupament sostenible

En 2015 l'ONU va aprovar una agenda d'objectius de desenvolupament sostenible per tal que els diferents països prengueren mesures per tal de millorar la societat. Aquesta agenda es pot classificar en una llista de 17 objectius fonamentals que involucra àmbits com la igualtat, l'eficiència energètica i industrial o protecció del medi ambient. A continuació presentem la relació que té el treball desenvolupat amb alguns dels punts d'aquesta llista d'objectius de desenvolupament sostenible:

- **Educació de qualitat** (Objectiu 4): La manca d'accés a la informació és un aspecte fonamental per a les carències educatives que existeixen en l'actualitat. La recerca en totes les àrees del PLN col·laboren en eliminar aquest problema, ja que faciliten l'accessibilitat a la informació. Amb la generació automàtica de resums, com hem comentat amb anterioritat, es millora la recerca d'informació abreujant els textos i presentant de forma clara la temàtica principal i el contingut dels documents.

Figura 1.2: Aquests són els 17 objectius de desenvolupament sostenible establerts per l'ONU a l'agenda per a 2030.

- **Reducció de les desigualtats** (Objectiu 10): Com presentem més endavant, una de les principals característiques d'aquest treball és la utilització d'un nou corpus de dades de gran dimensió en català i en castellà que suposa uns dels majors corpus de notícies en ambdues llengües. Com és ben sabut, el català és una llengua minoritzada que ha aplegat a ser censurada i actualment es troba en situació de desigualtat fins i tot als llocs on podem trobar la gran majoria de catalanoparlants. Amb la utilització i promoció d'aquest corpus incentivem la recerca en català i facilitem el desenvolupament de ferramentes basades en el PLN en aquesta llengua.

1.4 Estructura de la memòria

La memòria d'aquest treball es compon de sis capítols que presentem a continuació junt a una breu descripció del contingut de cadascuna de les seccions amb la finalitat de donar al lector una visió general del treball.

- **Capítol 1, Introducció.** En aquest capítol es descriu l'àmbit del present treball i quina és la motivació i el context pel qual es du a terme. També s'exposa quina és la principal finalitat del treball i els objectius que es persegueixen.

- **Capítol 2, Estat de l'art.** El segon capítol busca contextualitzar el treball i l'actualitat dels camps de recerca que es veuen implicats en la seua realització. Per a aquest fi es tracta l'evolució de la generació automàtica de textos, les actuals tendències i les noves tendències basades en xarxes neuronals presents a la generació automàtica de resums.

- **Capítol 3, Metodologia i sistemes utilitzats.** En el capítol tercer tractarem els sistemes de representació de textos, els sistemes de generació de resums i les mètriques que s'han fet servir. També s'explica la naturalesa de les dades que s'utilitzen per a la realització de l'estudi.

- **Capítol 4, Ferramentes utilitzades.** En aquest capítol es pretén exposar i explicar de manera detallada les ferramentes que han sigut necessàries per a la realització de les diferents tasques requerides per a l'experimentació.

- **Capítol 5, Experimentació i resultats.** Al cinqué capítol es descriu el seguiment de les labors realitzades i s'exposen els diferents resultats obtinguts a partir de les pautes descrites al capítol anterior. A continuació es comparen aquests resultats mitjançant l'aplicació de diferents mesures i es realitza una valoració de les solucions aplicades.

- **Capítol 6, Conclusions.** En aquest últim capítol exposem les idees finals extretes de l'obtenció de resultats i s'avalua l'assoliment dels diferents objectius. També es detallen els diferents obstacles que ha presentat la realització del treball i s'expliquen les diferents vies de treball futures per les que es pot continuar arran d'aquest treball.

1.5 Context i col·laboracions

El present treball s'ha desenvolupat durant el gaudiment d'una beca de col·laboració oferida pel Ministeri d'educació i de formació professional. Aquestes beques tenen la finalitat de que l'alumnat duga a terme tasques d'investigació en departaments universitaris.[3] En aquest cas particular s'ha realitzat en col·laboració amb el VRAIN[4] (*Valencian Research Institute for Artificial Intelligence*) de la Universitat Politècnica de València (UPV), més concretament en un dels seus set grups d'investigació, Enginyeria del Llenguatge Natural i Reconeixement de Formes (ELiRF).

Aquest grup porta a terme les seues activitats al Departament de Sistemes Informàtics i Computació (DSIC) de la UPV. Les seues línies d'investigació giren al voltant de les tecnologies de la parla, centrades en el reconeixement automàtic i la creació de sistemes de comprensió i diàleg parlat; el tractament del llenguatge natural, on engloben l'anàlisi de textos, la recuperació d'informació, anàlisi de sentiments en xarxes social i el desenvolupament de sistemes de generació automàtica de resums basats en xarxes neuronals [13].

Dins del mateix grup s'ha estat en contacte principalment amb Encarna Segarra Soriano, Lluís Felip Hurtado Oliver, José Ángel González Barba i Vicent Ahuir Esteve. El conjunt de dades amb les quals es treballa en aquesta investigació ha sigut proporcionat per ells, d'igual manera que han ajudat a organitzar les tasques, facilitat ferramentes desenvolupades en el mateix grup i han disposat recursos específics essencials en determinades labors, necessàries per a l'adequada realització del projecte.

[3] Per a més informació es pot consultar el següent enllaç: https://www.educacionyfp.gob.es/ca/servicios-al-ciudadano/catalogo/general/99/998142/ficha/998142-2021.html.

[4] Consultar el lloc https://vrain.upv.es/index.php per a saber més.

1.6 Relació amb els estudis cursats

Durant la realització del grau en enginyeria informàtica a la Universitat politècnica de València, més concretament al bloc de computació, s'han presentat diferents conceptes que han servit per a la comprensió i coneixement previ de diversos conceptes que s'han treballat en el present treball. En primer lloc, en l'assignatura d'algorítmica (ALT) ha servit per a l'optimització i el càlcul de costos dels algoritmes desenvolupats.[5] Altres coneixements sobre l'estructura bàsica de les xarxes neuronals i el funcionament i entrenament de sistemes d'aprenentatge profund s'han obtingut en assignatures com Percepció (PER) i Aprenentatge automàtic (APR) respectivament. Per últim, l'assignatura més destacada és la de Sistemes d'emmagatzematge i recuperació d'informació (SAR). En ella es presenten diferents sistemes de representació de documents que es tracten en la present memòria i diversos conceptes sobre el PLN.

[5] Ens referim, més concretament, a l'oracle desenvolupat en el present treball.

CAPÍTOL 2
Estat de l'art

En aquest capítol tractarem inicialment els diferents mètodes i vies d'estudi que estan presents en l'àrea d'investigació del PLN i, més concretament, en la generació automàtica de resums. A continuació, exposarem algunes de les tendències punteres en l'actualitat conjuntament amb les tècniques i metodologies més destacades o amb millors resultats. Finalment exposarem alguns exemples de treballs que són referents a l'hora de comparar els sistemes de generació automàtica de resums.

2.1 Processament del Llenguatge Natural

El processament del llenguatge natural és una àrea molt activa i en constant desenvolupament. En conseqüència la seua definició ha anat variant amb el transcurs del temps i compta amb múltiples interpretacions. Elizabeth D. Liddy en [8] ens ofereix la següent definició:

> *Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.*

D'aquesta definició es poden extraure diferents aspectes de les característiques que presenta el PLN. En primer lloc, esmentar que la frase *human-like language processing* permet identificar que es tracta d'una disciplina de la Intel·ligència Artificial. Com hem introduït anteriorment, hi ha diferents maneres d'intentar replicar el comportament humà, per exemple recreant el seu funcionament o aconseguint obtenir els mateixos resultats [36]. En aquesta branca no és diferent. Un dels objectius finals del PLN és que els ordinadors tinguen la capacitat d'entendre el llenguatge humà i, en conseqüència, poder complir les tasques que es requereixen amb el mateix nivell d'enteniment que les persones. Aquest fenomen se'l coneix com a NLU (*Natural Language Understanding*) [8].

Actualment aquesta meta no s'ha aconseguit, principalment per la complexitat del llenguatge humà, el nombre d'idiomes utilitzats al voltant del món i la quantitat de contexts diferents que es poden presentar. Fins ara la recerca ha anat encaminada majoritàriament a la resolució de tasques concretes com poden ser la traducció de textos, cerca de respostes o transcripció de discursos orals. En aquests casos s'han aconseguit molt bons resultats en distintes disciplines.

Tot i això, hui en dia s'han desenvolupat diferents sistemes dirigits a la resolució de múltiples tasques del PLN en una única ferramenta. Actualment destaquen GLUE [2],

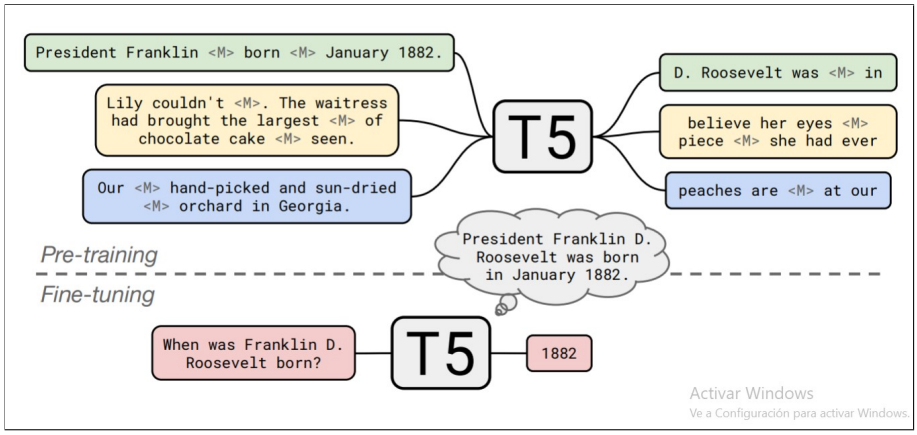

Figura 2.1: Representació visual de com el sistema de T5 serveix per a diverses tasques de PLN com traducció o cerca de respostes [1].

SuperGLUE [3] i T5 (*Text-to-Text Transfer Transformer*) [5]. Els models d'aquestes ferramentes tenen la finalitat d'aconseguir bons resultats en no totes però si en un gran nombre de tasques del PLN. Aquests sistemes utilitzen diferents tecnologies, però tots comparteixen l'ús de models d'aprenentatge profund. Aquesta tendència és la principal hui en dia, ja que, actualment i gràcies a les quantitats inabastables de text disponible a la *World Wide Web*, els resultats són molt més prometedors que amb sistemes probabilístics clàssics [32].

2.2 Generació automàtica de resums

A causa de l'augment de la disponibilitat de documents de text a la xarxa la importància de la generació de textos no ha fet més que augmentar. Tot i això, aquesta disciplina del PLN ha estat una de les principals des de les primeres etapes. Radev ens ofereix una idea prou clara del que es busca amb la generació de resums amb la següent definició:

> *A summary can be loosely defined as a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that [...] The main goal of a summary is to present the main ideas in a document in less space [6].*

Quan les persones generem un resum d'un text realitzem un gran nombre d'assumpcions que hem anat aprenent al llarg del temps, tant sobre el funcionament del llenguatge com de coneixement concret dels temes dels quals tracta el document o documents en qüestió. És per aquest motiu, ja que les màquines no disposen dels mateixos mecanismes d'aprenentatge que les persones, que la complexitat de generar resums de manera automàtica pot resultar molt més elevada del que es pot suposar en una primera instància [26].

Dins de la generació de resums cal considerar les diferents variants que pot seguir l'enfocament. Primerament es sol distingir entre si la generació del resum és extractiva o és abstractiva. Quan es tracta d'un resum extractiu parlem d'un resum del text constituït completament per material present en el text original. Per altra banda, si algunes parts rellevants[1] del resum no es poden trobar en el text que s'ha utilitzat per a generar

[1] Per "rellevant"ens referim a tot aquell contingut que aporte informació, és a dir, que no siguen signes de puntuació o *stopwords* que servisquen de connectors entre fragments de text per exemple.

el resum estarem parlant d'un resum abstractiu. En segon lloc, la generació de resums també es poden distingir per estar destinats a resumir un únic text o obtenir un resum general d'un conjunt de textos (*multi-document summarization*). Finalment els resums poden estar dirigits a un grup o tasca concreta (*user-focused* o *topic-focused*) o ser de caràcter general. Actualment, amb la tendència en la recerca del PLN de poder realitzar múltiples tècniques en un mateix sistema els resums *user-focused* estan guanyant una gran rellevància [12].

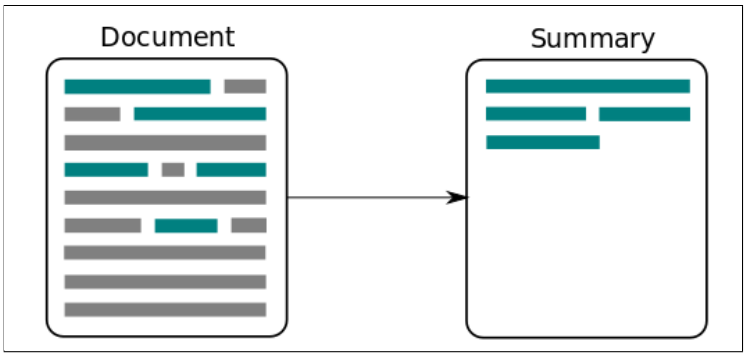

Figura 2.2: Exemple gràfic de la definició de resum. Ja que les frases que componen el resum de la il·lustració són distintes a les que es seleccionen del document original es pot utilitzar també per exemplificar un resum abstractiu. Per a tractar-se d'un resum extractiu caldria que el resum continguera els fragments originals del document.

Des de les primeres aproximacions ja es va plantejar la idea de compondre el resum de les millors frases després de puntuar-les mitjançant diferents sistemes. Els plantejaments inicials es van basar en puntuar les frases del text per la presència de paraules destacades del text, per compartir paraules amb el títol del document o basant-se en la idea que les primeres frases són aquelles que contenen la informació més rellevant. D'aquesta última modalitat cal destacar Lead-K, ja que es farà servir en el desenvolupament del projecte. Lead-K és una heurística que consisteix a seleccionar les K primeres frases d'un text, de vegades amb limitacions per al nombre de paraules. Aquesta tècnica ha tingut comunament una gran efectivitat en textos de noticiaris.

En l'actualitat, les aproximacions basades en models coneguts com a *híbrids* són de gran interés. Aquests models incorporen la generació automàtica de resums de manera que al procés d'extracció que hem explicat anteriorment se li suma la incorporació ,ja siga de manera complementària o conjunta, de la generació abstractiva. Com a resultat, aquestes tècniques estan aconseguint resums concisos i intel·ligibles amb millors resultats que les tècniques extractives tradicionals. Dues de les propostes d'aquesta aproximació que presenten resultats punters en l'àrea de treball i que representen les dues tendències presents dins de la mateixa són el *Reinforcement Learning model* (RL) [40] i el *Inconsistency Loss model* (IL) [39].

El model RL es basa en l'estratègia *extraccion-then-abstraccion*. En aquest model es realitza un entrenament de reforç per enllaçar el funcionament dels dos mòduls per separat. L'entrenament d'aquests models es realitza amb documents etiquetats[2] de manera que la part extractiva recupera la frase del document amb major semblança[3] i l'etiqueta com positiva i la resta com negatives (cosa que permet reformular aquesta tasca com un pro-

[2]En el context de la generació de resums ens referim per etiqueta a un resum real o generat amb unes finalitats o condicions específiques per tal que els sistemes puguen conéixer com hauria de ser o a què s'hauria d'assemblar el resum resultant.

[3]La semblança ve determinada per la funció de mesura que s'utilitze. En aquest cas, com és habitual en la comparació de semblança entre textos, es fa servir la mesura ROUGE-L, més concretament el seu valor de *recall*. Sobre aquesta mesura es parlara de manera més elaborada en capítols posteriors de la memòria.

blema de classificació). A partir d'aquesta frase la part abstractiva intenta comprimir el fragment perquè aconseguisca una major similitud amb el resum original[4].

Per la seua banda, el model IL presenta l'estratègia *extraccion-with-abstraccion* i aconsegueix una millor llegibilitat.[5] Tot i que l'estructura de l'entrenament és similar al model RL, aquesta tècnica posa el focus en combinar una atenció[6] en l'àmbit de oració en la part extractiva i l'atenció en l'àmbit de paraula en la part abstractiva. Finalment, amb la finalitat que les dues atencions siguen coherent fan servir la funció de *inconsistency loss*, el que assegura que les frases amb una alta atenció la tenen també quant a paraula [22].

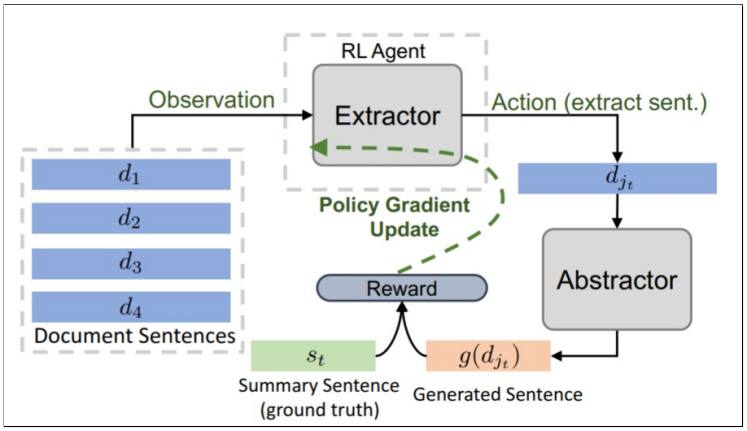

Figura 2.3: Funcionament d'aprenentatge i puntuació del model RL [40].

Per altra banda si que trobem sistemes de generació automàtica de resums abstractius. Aquests sistemes seguixen presentant una fase de selecció de les parts més rellevants del text original i generen un resum parafrasejant o reformulant frases del document. Un exemple d'aquest tipus de sistema és PEGASUS.

Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence to sequence models o PEGASUS és un mètode de generació de resums abstractius que ha obtingut resultats destacables en diverses feines. Aquests resums es generen a partir de frases del document seleccionades segons el que ha aprés el sistema mitjançant entrenament. Amb la finalitat de generar resums sense necessitar l'ús d'una part extractiva opta per aplicar la tècnica de *Gap Sentence Generation*. Aquesta tècnica consisteix a seleccionar algunes frases del document (basant-se en mètodes com lead-K, aleatori o les que aconseguisquen les millors puntuacions utilitzant alguna mètrica).

Aquest sistema ha sigut preentrenat amb múltiples corpus i a l'hora de ser avaluat ha demostrat que és capaç d'obtenir bons resultats en avaluació humana. A més a més, per mitjà d'experimentació s'ha pogut observar que és capaç d'adaptar-se amb facilitats a conjunts de dades amb els que no ha treballat anteriorment. PEGASUS ens permet concloure que la situació de la generació automàtica de resums abstractius és prometedora i dóna peu a la possibilitat d'obtenir millors resultats que els obtinguts fins ara amb les tècniques habituals [16].

Ara com ara, la popularitat i activitat en investigació d'aquest sector ha crescut acceleradament gràcies, entre altres coses, a l'èxit que estan tenint noves tècniques d'aprenentatge i aplicacions de sistemes d'intel·ligència artificial. Des de la presentació del perceptró per Rosenblatt [9], l'aprenentatge automàtic ha mostrat la gran capacitat que té

[4] L'etiqueta associada al text amb el qual treballem.

[5] Com veurem més endavant, fem referència a la mesura ROUGE-2, ja que segons alguns estudis és la que permet mesurar millor la llegibilitat del text generat.

[6] Forma de mesurar la importància d'un fragment. Sol dependre de la mètrica utilitzada o la funció a maximitzar.

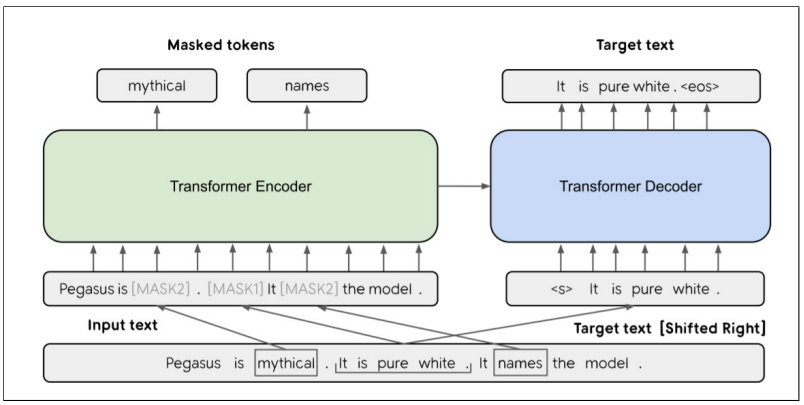

Figura 2.4: Arquitectura de PEGASUS [16].

per a resoldre problemes del món real i el potencial per a contribuir al desenvolupament de diverses disciplines. A més a més, l'aprenentatge profund o *deep learning*, una branca de l'aprenentatge automàtic, està guanyat transcendència a l'haver demostrat que pot aplegar a ser de gran utilitat per a millorar resultats en algunes tasques o permetre dur a terme noves interpretacions de problemes estudiats únicament en l'àmbit teòric per a la seua resolució.

L'aprenentatge profund es pot simplificar com l'aplicació de xarxes neuronals amb múltiples capes de neurones (4 o més habitualment) entre l'entrada i l'eixida. La importància del *deep learning* recau en la seua capacitat de representar relacions no lineals entre objectes i representar-ho d'una manera molt menys abstracta. A més és capaç d'aprendre funcions extremadament complexes sense l'especificació per part d'un humà. És a dir, gràcies als mecanismes d'aprenentatge que presenten aquests sistemes, són capaços de generalitzar aprenent de manera autònoma [34].

Aquesta disciplina no s'aplica únicament a la generació automàtica de resums ni al PLN, sinó que engloba pràcticament totes les disciplines de la intel·ligència artificial. En alguns casos certes tecnologies de l'aprenentatge profund s'especialitzen en àrees concretes de la intel·ligència artificial a causa de les característiques i requeriments. Un d'aquests casos és el de les xarxes neuronals convolucionals[7] (CNN)[8] les quals destaquen per la seua presència en l'estudi del tractament d'imatges i la detecció d'objectes. No obstant això, en l'actualitat la investigació en el PLN ha trobat la manera d'incorporar aquesta tecnologia en les seues diverses àrees [25].

Com hem vist anteriorment, en el camp de la generació automàtica de resums hi ha tecnologies punteres que basen els seus models en xarxes neuronals i en sistemes d'aprenentatge profund. També cal destacar la tendència actual en el PLN d'aplicar sistemes basats en xarxes neuronals coneguda com a *transfer learning*.

La tendència esmentada consisteix a utilitzar característiques o parts de xarxes neuronals que han sigut entrenades per a una determinada tasca amb la finalitat que siga aprofitable per a una altra distinta. Normalment el nexe que relaciona els dos sistemes sol ser o el conjunt de dades que s'utilitza per a realitzar l'entrenament o una gran similitud en-

[7] Les xarxes neuronals convolucionals són un tipus de xarxa neuronal composta per múltiples capes amb un gran nombre de neurones que progressivament va reduint el nombre de neurones per capa per tal de traure un resultat més abastable, per exemple tornant un valor de probabilitat de pertinència a una classe en tasques de classificació. En les diferents capes fa aprenent a extraure característiques cada vegada més complexes de les entrades que rep per a finalment poder realitzar la labor que es necessita. Es caracteritza també per reduir cada cert nombre de capes mitjançant sistemes de filtratge o selecció com pot ser *max pooling* [29]. Per a conéixer més d'aquest tipus d'arquitectura es recomana consultar [35].

[8] *Convolutional Neural Networks*.

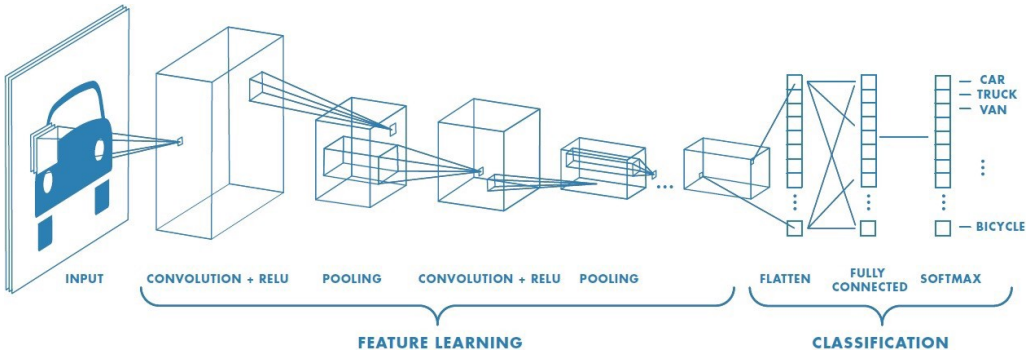

Figura 2.5: Representació de l'arquitectura d'una xarxa neuronal convolucional.

tre les tasques. Un dels principals avantatges és que permet evitar el sobreentrenament[9] i disminuir els temps d'entrenament [23].

Altrament, una tecnologia que es pot trobar en un gran nombre d'arquitectures punteres en matèria de generació automàtica de resums, i en la recerca del PLN en general, són els *transformers* [18]. Els *transformers* són un model de *Deep Learning* que utilitza un mecanisme de *self-atention*. Recolzant-se en un conjunt de vectors d'atenció determina per a cada quines paraules del context són rellevants.

Quan se li dóna una frase a un *transformer*, l'atenció de cada token es calcula de manera simultània (ja que aquesta no depén d'haver processat primer les paraules anteriors) i s'obtín una representació mitjançant *embeddings* per cada paraula amb informació de la pròpia paraula i que varia segons la influència calculada de l'aportació de la resta de tokens. Alguns dels sistemes importants que presenten aquests models són BERT, PEGASUS o BART.

Altres adaptacions entorn de l'aprenentatge automàtic també s'han traslladat en major o menor mesura. Entre elles podem trobar xarxes neuronals, màquines de vectors de suport [21], o SVM (*Support Vectors Machine*), xarxes neuronals recurrents, o RNN (*Recurrent Neural Network*) i xarxes neuronals siameses [19] de les quals tractarem en major profunditat en el següent capítol, ja que és una de les tecnologies utilitzades per a la generació de resums en aquest treball.

2.3 Corpus

Enfront de la tendència de l'ús de models d'aprenentatge profund la generació automàtica de resums els conjunts de dades d'alta qualitat s'han tornat fonamentals. Aquestes dades són essencials per als processos de preentrenament per a obtenir valors inicials des dels quals partir per a especialitzar-se en tasques concretes. També l'ajust de paràmetres d'aquest tipus d'arquitectures requereix corpus de dades consistents en parells de documents i resums. L'adquisició d'aquest tipus de corpus no és una feina trivial, ja que requereix una gran participació humana a l'hora de crear els resums o buscar noves maneres d'obtenir aquests resums de manera automatitzada.

En l'actualitat, la gran majoria de ferramentes i sistemes estan desenvolupats en i per a l'anglés, és a dir, han sigut entrenats i avaluats en textos en llengua anglesa i resulten d'utilitat sobretot per a aquest idioma. La major part de les eines de generació automàtica

[9] Entenem per sobreentrenament al sobreajust en el procés d'aprenentatge d'una xarxa neuronal, resultant en conseqüència en una perduda de la genericitat. Arran d'açò els resultats obtinguts i la utilitat de la xarxa entrenada perd utilitat i validesa a l'hora de ser utilitzada en entorns diferents que en el que s'ha desenvolupat.

de resums treballa, entre altres conjunts, amb els *datasets* CNN/Daily Mail i NewsRoom. El conjunt de dades CNN/Daily Mail és un corpus d'articles de notícies en llengua anglés compost per prop de 300 000 parells únics de notícies i resums procedents dels diaris CNN i *Daily Mail*.

Per altra banda, NewsRoom és un conjunt d'1,3 milions d'articles i notícies escrits pels respectius autors i editors de 38 mitjans de comunicació extrets directament dels llocs webs dels propis mitjans. Partint de la presentació d'aquest corpus i la metodologia d'obtenció s'ha creat el corpus DACSA amb el que treballarem en aquest treball i que explicarem en el següent capítol.

CAPÍTOL 3
Metodologia i sistemes utilitzats

3.1 Descripció del corpus

Com hem avançat en els anteriors capítols de la memòria, en el present treball el conjunt de dades que utilitzarem són les extretes del corpus DACSA. Aquest corpus està compost per notícies de diferents mitjans de comunicació les quals tenen associats resums creats pels autors o editors de les notícies. La creació d'aquest conjunt de dades s'ha realitzat seguint un procediment similar al de NewsRoom, és a dir, s'ha extret les notícies de les pàgines web periodístiques dels mitjans de comunicació. D'aquesta manera s'ha aconseguit un conjunt de dades de quasi 1 000 000 de notícies per al català i prop de cinc 5 500 000 de notícies per a castellà. La distribució per fonts queda de la següent manera:

Font Periodistica	**Documents**	**Exclosos**	**Inclosos**
Diari ARA	288 081	49 848	238 233
Diari de Griona	224 705	87 258	137 447
Diari La Veu	49 494	13 731	35 763
El Per. de Catalunya	234 022	39 325	194 697
El Punt Avui	10 170	3 066	7 104
El Temps	6 887	1 005	5 882
Nació Digital	56 110	11 729	44 381
Regió 7	64 180	7 353	56 827
VilaWeb	25 663	20 813	4 850
Total	959 312	234 128	725 184

Taula 3.1: En aquesta taula es mostra com es distribueix la procedència de les notícies i resums en català per a la creació del corpus DACSA. També es distingeix el nombre de notícies que han sigut o no excloses de les particions del corpus.

Per a crear DACSA s'ha aplicat sobre aquestes notícies una sèrie de filtres per a assegurar que els documents compliren una sèrie de característiques com tindre un nombre mínim de paraules a l'article i al resum o no aplegar a una cota màxima de similitud entre el començament de la notícia i el resum. Els parells que no han complit alguns d'aquests filtres han sigut exclosos. Com a resultat han quedat inclosos en les particions de DACSA uns 725 000 articles per al conjunt en català i més de 2 100 000 per al castellà.

Les notícies que han quedat incloses han sigut separades en 4 particions: **train**; destinat a entrenar models i sistemes de PLN, **validation**, per tal d'acompanyar al conjunt d'entrenament en sistemes que utilitzen grups de validació (el qual és el comportament habitual per tal d'assegurar un bon entrenament); **test**, conjunt de notícies per a provar el funcionament de sistemes de tractament de textos; i **hard test**, Aquesta partició és un

Font Periodistica	Documents	Exclosos	Inclosos
20 Minutos	2 047 404	1 960 950	86 454
ABC	220 806	52 741	168 065
Diario de Mallorca	256 513	60 103	196 410
Agencia EFE	17 747	368	17 379
El Diario	127 618	11 057	116 561
El Economísta	178 026	176 652	1 374
El Español	157 092	9 039	148 053
El Independiente	42 018	25 053	16 965
El Mundo	112 592	77 480	35 112
El País	61 3954	63 806	550 148
El Per. de Aragón	60 605	3 370	57 235
El Per. de Catalunya	82 092	8 068	74 024
Europa Press	470 987	470 344	643
Expansión	73 395	73 240	155
La Razón	96 210	61 047	35 163
Las Provincias	598 346	256 301	342 045
La Vanguardia	129 581	30 483	99 098
Levante	86 594	4 647	81 947
Nodo 50	2 614	164	2 450
Público	119 770	12 608	107 162
Última Hora	4 100	3 633	467
Total	5 498 064	3 361 154	2 136 910

Taula 3.2: En aquesta taula es mostra com es distribueix la procedència de les notícies i resums en castellà per a la creació del corpus DACSA. També es distingeix el nombre de notícies que han sigut o no excloses de les particions del corpus.

conjunt distint de fonts que no es veuen en cap etapa de l'entrenament ni el testatge. Les fonts que s'han separat ha sigut amb motiu del baix nombre de notícies que preentaven. D'haver sigut incloses junt amb la resta de fonts seria probable que es donaren resultats no desitjats a l'hora d'avaluar o que al model a entrenar li costara més entrenar en centrar-se en l'estructura de les fonts majoritàries. La resta de fonts ha sigut separada en proporció 90% per al conjunt d'entrenament, 5% per al conjunt de validació i 5% per al conjunt de testeig cada una. La distribució per idioma d'aquestes particions queda com podem observar a la taula 3.3.

Idioma	Train	Validation	Test	Hard Test
Català	636 596	35 376	35 376	17 836
Castellà	1 802 919	104 052	104 052	109 626

Taula 3.3: Distribució del conjunt de notícies en les seues 4 particions per a cada una de les llengües de les quals disposa el corpus.

Com hem comentat al segon capítol la gran majoria de models dirigits al PLN estan entrenats en anglés o van dirigits principalment a realització de tasques en relació amb aquesta llengua. Per altra banda, sí que hi ha sistemes que funcionen en múltiples llengües incloent el castellà i el català. Tot i això, els resultats que presenten en aquestes llengües concretes no tenen el mateix nivell de rellevància, ja que aquestes recerques no apleguen a ser tan destacades o es disposa de menys recursos per a l'entrenament de models. En el cas del català es sol recórrer a contingut obtingut de la Viquipèdia o al corpus d'Oscar [30], *dataset* que fa servir BART per al seu entrenament i el qual és relativament reduït.

Tenint aquests fets en consideració podem destacar de manera més clara la importància del corpus de notícies que treballem en aquest projecte, ja que suposa un corpus de gran mida que sobrepassa a la majoria conjunts de notícies disponibles en l'àmbit general.

3.2 Representació de textos

Com hem comentat amb anterioritat, la representació de les dades és un aspecte fonamental en l'aprenentatge de models per a la realització de diferents tasques. A continuació mostrarem algunes de les representacions més importants i introduirem el concepte d'*embeddings*, el qual presenta una gran importància en aquest treball i en gran part de les tecnologies punteres actuals.

3.2.1. One-Hot

El primer sistema de representació que explicarem és la codificació one-hot. Aquesta representació consisteix a escriure cada paraula com un vector de talla n tal que $x = \{x_1, x_2, ..., x_n\}$, sent n la talla del vocabulari, presentant un 0 en les posicions on la paraula no es correspon amb la paraula del vocabulari i un 1 on sí que coincideix. De manera formal la representació d'una paraula es pot descriure com [17]:

$$x(i) = \{x_1, x_2, ..., x_n\} : x_i = 1, x_j = 0 \forall j \neq i$$

La representació d'un text en aquest sistema resulta en una matriu de dimensions $m * n$ on m és el nombre de paraules del text. Podem observar a la matriu de la figura 3.1 com, per exemple, la representació per a la paraula *ring* en aquest context seria $\{0, 0, 0, 0, 0, 1\}$.

	the	king	put	on	the	ring
the	1	0	0	0	1	0
king	0	1	0	0	0	0
put	0	0	1	0	0	0
on	0	0	0	1	0	0
ring	0	0	0	0	0	1

Figura 3.1: Exemple de la representació d'una frase amb la codificació One-Hot.

Els principals avantatges que presenta aquest sistema és la facilitat per a ser implementat i per a localitzar i eliminar errors. Per altra banda el principal problema és la ineficient quantitat d'espai requerit, ja que genera matrius enormes que, en ser molt disperses, emmagatzemen poca informació. A més a més, aquest tipus de representació l'informació que presenta sobre l'ordre de les paraules obliga a mantenir una dimensionalitat fixa i no aporta informació sobre la semàntica del document.

3.2.2. Bossa de paraules

Una altra representació de documents molt popular és la Bossa de paraules, també coneguda com *Bag of words*. En aquest mètode de representació es genera un vector de valors

per a cada paraula del vocabulari d'un document. La representació més simple que es pot generar amb la Bossa de paraules és la d'un vector que indique el nombre d'aparicions d'una paraula en el document. Aquest vector es podria generar fàcilment a partir de la codificació One-Hot mitjançant un sumatori de les files de la matriu. Una altra alternativa seria la compressió de la representació One-Hot mitjançant aplicar una funció or sobre les files. D'aquesta manera es representaria l'absència o presència de paraules del vocabulari en un únic vector.

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Figura 3.2: Representació de diferents frases mitjançant la Bossa de paraules.

A partir d'aquesta representació i utilitzant un conjunt de documents trobem la representació tf*idf (*Term Frequency*Inverse Document Frequency*). Aquesta representació es calcula a partir dues mesures basades en el nombre d'aparicions de les paraules (t) dins del document (d) i segons el nombre de documents en el conjunt(D) i la seua presència en els documents. A continuació presentem les fórmules (sotmeses a canvi segons la finalitat per a la qual s'utilitzen) per al càlcul dels valors de la representació:

$$tf(t,d) = \frac{f(t,d)}{\sum f(t',d) : t' \in d}$$

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}}$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

Amb aquesta representació hem vist que, tot i que l'ordre d'aparició es perd completament, sí que guanya una major importància la freqüència relativa de les paraules i el seu nombre d'aparicions en els documents. A més a més, el cost espacial d'aquesta representació és molt més eficient que la que presenta One-Hot. Una possible solució per al problema de l'ordenació seria realitzar el càlcul amb n-grames[1] [4].

3.2.3. Embeddings

Els *embeddings* són la representació que més atenció està rebent actualment en l'àrea del PLN. Els *emeddings*, a diferència de les dues representacions anteriors basades en representacions vectorials discretes, plasma millor, mitjançant vectors, les relacions semàntiques entre paraules mitjançant representacions distribuïdes basades en xarxes neuronals. Altrament, aquesta representació és significativament més eficient, ja que es realitza mitjançant un vector per a cadascuna de les paraules del vocabulari i d'una longitud fixa relativament reduïda.

[1] Conjunt de n paraules consecutives a les frases dels documents.

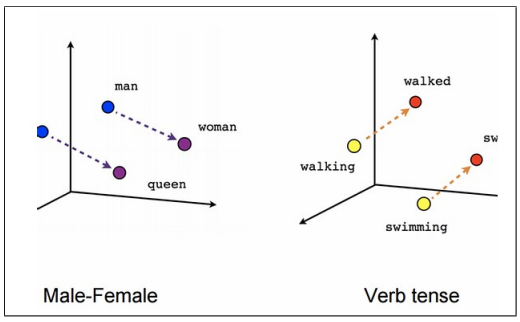

Figura 3.3: S'observa com parelles de paraules relacionades semànticament de la mateixa manera presenten una disposició similar.

La part interessant d'aquesta representació és que els vectors guarden informació d'alguns aspectes implícits del llenguatge com poden ser el temps verbal i el gènere dels substantius, com es mostra a la figura 3.3 o relacions entre països com a la figura 3.4. Per exemple, el resultat del càlcul $vec("Madrid") - vec("Spain") + vec("France")$ estarà més prop a vector de la paraula "París"que al de qualsevol altra, ja que aquests vectors permeten realitzar aquest tipus d'operacions senzilles amb resultats coherents [37].

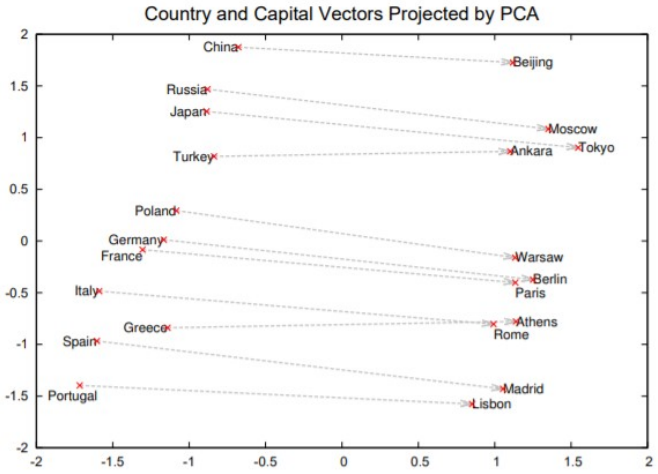

Figura 3.4: Representació mitjançant pca de la distribució en *embeddings* dels noms de països i les seues capitals [37].

A continuació descriurem les dues agrupacions principals que es poden distingir dins d'aquesta representació: els *embeddings* incontextuals i els *embeddings* contextuals.

Embeddings incontextuals

Els *embeddings* incontextuals són aquells que únicament presenten un vector per paraula. El propòsit és el de poder generalitzar les probabilitats de que aquesta paraula estiga acompanyada d'altres tenint en compte tots els contextos alhora. Alguns dels models més utilitzats per als *embeddings* incontextuals són Word2Vec i FastText.[2]

Per als models Word2Vec, per als quals cada paraula té assignat un vector, s'usa el model d'entrenament proposat per Mikolov, et al [38]. La finalitat de l'arquitectura *Skip-gram* es la de predir el context d'aparició d'una paraula a partir de maximitzar la log probabilitat de la seua presència segons les paraules del seu entorn. La definició formal de la maximització que realitza és la següent:

[2]Per trobar més informació sobre aquests models es recomana la referència [31].

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} logp(w_{t+j}|w_t)$$

On la paraula sobre la qual calcular el vector és w_t i c la mida del context a calcular. La fórmula bàsica de *skip-gram* per al càlcul de $p(w_{t+j}|w_t)$ utilitzant la funció s*softmax* es defineix com:

$$p(w_O|w_I) = \frac{exp(v'_{w_O}{}^{\top} v_{w_I})}{\sum_{w=1}^{W} exp(v'_w{}^{\top} v_{w_I})}$$

Però aquest càlcul no és eficient perquè el cost de calcular $logp(w_O|w_I)$ és proporcional la grandària del vocabulari, el qual sol ser molt gran. En el seu lloc es sol optar per *Hierarchical softmax*, una aproximació del càlcul de Softmax que únicament necessita realitzar $log_2 W$ avaluacions per al càlcul de la distribució de probabilitats [10].

Embeddings contextuals

Altrament els *embeddings* contextuals distingeixen els pesos associats a cada paraula segons el context. Cada paraula del vocabulari té associat un vector de pesos que varia depenent de les circumstàncies en les quals la paraula apareix. Un exemple d'arquitectura que gasta aquesta representació és BERT [14]. BERT (*Bidirectional Encoder Representations from Transformers*) és un sistema que, a partir de text no etiquetat i centrant-se en el context a dretes i a esquerres de cada paraula, aprén representacions contextuals i bidireccionals. BERT primerament fa servir el model de llenguatge emmascarat (*Masked Language Model*). Aquest model consisteix a ocultar certes paraules de les frases i intentar reomplir els espais correctament. Al tractar-se d'un sistema bidireccional no només emmascara per predir d'esquerra a dreta sinó que també de l'inrevés.

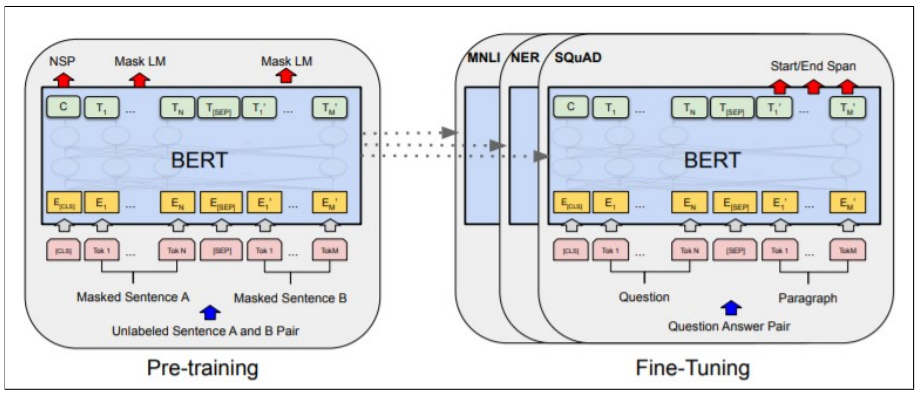

Figura 3.5: Representació de les etapes de *pre-training* i *fine-tuning* que realitza BERT per a l'aprenentatge dels seus paràmetres [14].

A continuació fa servir la tècnica *next sentence prediction* per tal de millorar la representació conjunta de fragments de texts. Aquestes dues etapes formen part del preentrenament (*pre-training*). A partir dels resultats obtinguts en la fase anterior s'inicia l'etapa d'ajust (*fine-tuning*). En aquest procés, s'utilitzen dades etiquetades per a millorar les prestacions dels paràmetres obtinguts mitjançant diferents tasques, cada una amb diferents models per a constituir posteriorment l'arquitectura final. En la figura 3.5 veiem un exemple d'aquest esquema.

3.3 Mètriques d'avaluació

Per tal de poder comparar els resums que generem en aquest treball amb els resums de referència de les notícies necessitarem fer ús de diferents mètriques. Amb aquestes mètriques es podran observar i avaluar les característiques dels diferents resums generats. Amb aquest propòsit s'han seleccionat la mètrica de ROUGE, per tal d'avaluar en l'àmbit sintàctic, i BERTScore per a obtenir resultats sobre la similitud dirigida més a la semàntica.

Cap de les dues mesures ha sigut implementada durant el desenvolupament del present treball, ja que es tracta de mesures utilitzades comunament per a aquest propòsit i podem trobar un gran ventall de ferramentes que inclouen la implementació d'aquestes mesures de manera optimitzada i de fàcil accés mitjançant llibreries. En el nostre cas hem optat per utilitzar les mètriques proporcionades per *Hugging Face* en *datasets.metrics*.

3.3.1. ROUGE

La mesura de ROUGE, la qual torna un valor entre 0 i 1, obtén el grau de similitud entre dos textos d'acord amb el solapament que hi ha entre ells. El valor de ROUGE només serà d'1 si els textos que es comparen són el mateix. En ROUGE podem distingir diferents aproximacions com són ROUGE-N, que mesura el solapament entre N-grames, i ROUGE-L, que busca la cadena coincident més llarga per a obtenir les puntuacions. A continuació explicarem aquestes aproximacions i les diferents versions que poden presentar i que utilitzem en aquest treball.

ROUGE-1

És la versió de ROUGE-N que obté el càlcul de la superposició de paraules individualment. Consisteix a calcular la superposició entre el nombre de paraules del resum generat i el resum de referència. Aprofitant que es tracta del cas més senill i més clar de calcular explicarem tres valors que es poden extraure de cada una de les mesures de ROUGE: *recall*, *precisió* i *f-score*. El valor de *recall* es pot interpretar com el nombre de paraules del resum de referència que es poden trobar en el resum que es compara (en el cas de ROUGE-1). Podem expressar la fórmula d'allò que descrivim com:

$$Recall = \frac{\sum_{t \in \{S_r\}} Count(S_g, t)}{|S_r|}$$

On S_r és el resum de referència, S_g és el resum de generat, $\{S_r\}$ és el vocabulari del resum de referència, $|S_r|$ és la talla del resum de referència en nombre de paraules i $Count(text, t)$ és una funció que retorna el nombre d'elements coincidents entre el nombre d'aparicions del token t en el text introduït i el nombre d'ocurrències de t en el resum de referència.

Seguidament presentem un exemple pràctic de l'aplicació d'aquesta fórmula considerant S_g = "*the cat was found under the bed*ï S_r = "*the cat was under the bed*"Per a aquest cas observem que totes les paraules del resum de referència es troben al resum generat. Per tant el resultat queda com $Recall = \frac{6}{6} = 1,0$.

La precisió es pot definir com el nombre de paraules del resum generat que són rellevants, és a dir, que apareixen en el resum de referència. La fórmula per a obtenir aquest valor és la següent:

$$Precisio = \frac{\sum_{t \in \{S_r\}} Count(S_g, t)}{|S_g|}$$

On es segueix la nomenclatura de la fórmula anterior. El valor de *precisió* serveix per a valorar la quantitat de text generat que no aporta informació present al resum de referència (de manera literal).

Seguint també amb l'exemple presentat en l'explicació del valor de *recall* obtenim que la talla del resum generat és de ser paraules. Com que només sis d'eixes set paraules es troben també al resum de referència obtenim que el valor per a la *precisió* és: $Precisio = \frac{6}{7} = 0,86$.

Finalment trobem el valor *f-score*. Aquest valor consisteix a combinar els dos valors anteriors aplicant, si cal, un escalar que done major importància al valor desitjat. La mesura *f-score* s'obtén de la següent manera:

$$F - score = \frac{(1 + \beta^2) R(S_g, S_r) P(S_g, S_r)}{R(S_g, S_r) + \beta^2 P(S_g, S_r)}$$

On la funció $R(S_g, S_r)$ obtén el valor de recall entre el resum generat i el de referència, la funció $P(S_g, S_r)$ torna el valor precisió entre el resum generat i el resum de referència, i β controla la importància relativa de les mesures.

En el cas del present treball no ens hem centrat en la cerca de resums que continguen únicament les paraules que es troben al resum de referència ni en generar resums que eviten tindre paraules que no es troben en els resums associats a les notícies. Per aquest motiu hem fet servir en tots els casos la mesura *f-score*, que ens ofereix una visió més general.

ROUGE-2

ROUGE-2 es comporta de la mateixa manera que ROUGE-1 amb la diferència que treballa amb bigames en lloc d'amb unigrames. Per als càlculs en lloc de per cada paraula es realitzen els sumatoris per parells de paraules i es genera un vocabulari de bigrames en lloc que de paraules individuals. Aquesta mesura té una gran importància en la generació de resums automàtics, ja que s'ha demostrat que està molt relacionada en la llegibilitat del text generat.

ROUGE-L

La mesura de ROUGE-L es distingeix de les anteriors en el fet que no realitza operacions sobre la superposició de n-grames (ROUGE-N), sinó que realitza el càlcul a partir de la cadena coincident de major longitud. En molts casos es considera que aquesta mesura és la més fiable per poder determinar la similitud entre el contingut entre els resums comparats.

ROUGE-Lsum

ROUGE-Lsum és una implementació alternativa de ROUGE-L que, en lloc de considerar els dos fragments de text que se li proporcionen com una unitat, separa el resum de referència en frases. Realitzant aquesta separació troba la cadena de major longitud coincident per a cada una de les frases del resum de referència i realitza el càlcul dels valors

a partir de les seqüències de tokens seleccionades. Les fórmules necessàries, considerant $LCS_{frase}(S_g, s)$ una funció que torna la subseqüència coincident més llarga (*Longuest Common Subsequence*) entre el resum generat S_g i la frase s, són:

$$LCS(S_g, S_r) = \cup_{r_i \in S_r} LCS_{frase}(S_g, r_i)]$$

$$Recall = R_{RL}(S_g, S_r) = \frac{|LCS(S_g, S_r)|}{|S_r|}$$

$$Precisio = P_{RL}(S_g, S_r) = \frac{|LCS(S_g, S_r)|}{|S_g|}$$

$$F - score_{RL}(S_g, S_r) = \frac{(1+\beta^2) R_{RL}(S_g, S_r) P_{RL}(S_g, S_r)}{R_{RL}(S_g, S_r) + \beta^2 P_{RL}(S_g, S_r)}$$

Aquesta és una implementació molt convenient per als resultats que es busca obtenir en l'experimentació i amb el comportament de l'oracle desenvolupat (del qual parlarem en aquest capítol), és per aquest motiu que li donarem una gran rellevància a l'hora d'avaluar els resums generats.

3.3.2. BERTScore

BERTScore és un mètode automàtic d'avaluació per a la generació de text. El seu funcionament es basa a obtenir un valor de similitud per a cada token del text generat sobre els tokens del resum de referència. Per a calcular aquesta similitud s'utilitzen *embeddings* contextuals preentrenats amb BERT. També es pot definir la mesura BERTScore com el sumatori de les similituds cosinus entre els *embeddings* dels tokens de les frases a comparar.

A diferència de ROUGE el rang de valors habituals no van de 0 fins a 1 indicant d'una manera lineal el grau de similitud entre les frases. En el cas de BERTScore és molt senzill obtenir valors superiors a 65 i molt difícil obtenir valors superiors a 85, almenys en idiomes com el català i el castellà. És per aquest motiu que la mesura BERTScore no ens permet interpretar exactament el resultat com el grau de similitud semàntica dels textos comparats, però sí que ens ajuda a establir un ordre entre resultats. És a dir, si el valor de similitud amb BERTScore entre un parell de resums és major al d'altre parell de resums es pot assegurar que el primer parell té major similitud que el segon parell encara que no siga fàcil determinar la magnitud de la diferència.

Per exemplificar la particularitat esmentada sobre el rang de valors que s'obtenen en les avaluacions amb aquesta mesura ens disposem a presentar dos casos extrems. En el primer d'ells executem la instrucció:

```
bert-score --lang ca -r "Cotxe Divendres quan" -c "No passat"
```

Com podem observar es tracta de dues frases sense cap relació observable i amb poca correcció semàntica. No obstat això, la puntuació que obtenim en la mesura és de $0,655962$. En el segon exemple busquem el contrari, comparar frases amb una estructura semblant i que expressen la mateixa informació:

```
bert-score --lang ca -r "Quan major és la longitud millor es puntua
la similitud" -c "Com més gran és la llargària major és la similitud"
```

Tot i això la puntuació que obtenim d'aquesta comparació és de $0,841035$. Per a superar aquests valors les frases han de ser extremadament semblants o idèntiques.

3.4 Sistemes de resum tradicional no supervisat

Ja que el corpus DACSA encara no ha sigut publicat no existeixen experiments sobre ell que permitisquen comparar els resultats amb altres models basats en xarxes neuronals. Per aquest motiu hem optat per utilitzar alguns sistemes de generació automàtica de resums no supervisats. Els tres sistemes tradicionals que hem elegit són Lead-K, TextRank i la creació d'un oracle.

3.4.1. Lead-K

La generació de resums mitjançant Lead-K consisteix a seleccionar les primeres K frases del text original. Aquestes K frases sense modificar formaran íntegrament el resum. La versió més popular és Lead-3, com hem vist que s'utilitza en el cas de PEGASUS. Tot i això en el nostre cas, a causa de que, com veurem més endavant en l'estudi del corpus, la mitjana de la longitud en frases dels resums de referència és d'entre 1 i 2, farem ús de Lead-2.

Aquest sistema és molt popular pel baix cost de recursos i de temps d'execució que suposa. També cal destacar que té una rellevància major en la generació de resums de notícies, ja que aquestes tendeixen a presentar la informació més important de la notícia al començament. Aquesta distribució pot resultar un problema, ja que hi ha mitjans de comunicació en què el resum associat a la notícia es compon únicament per les 3 primeres frases d'aquesta. Per aquest motiu serà convenient la realització d'un estudi del corpus que permeta comprovar si el problema comentat passa en DACSA.

3.4.2. TextRank

TextRank és un model basat en grafs per a l'ordenació de fragments de text per importància. Serveix tant com per a trobar les frases més rellevants com per a trobar les paraules clau d'un document. Per tal de construir el graf per a la puntuació de frases TextRank genera un vèrtex per cada una d'elles i assigna un pes a les arestes igual al nombre de paraules que es superposen entre les frases que s'uneixen. Per a determinar les paraules clau genera un graf amb les paraules del text i enllaça aquelles que han aparegut juntes al text. A mesura que es repeteix aquesta contigüitat el pes d'aquesta aresta va incrementant[33].

3.4.3. Oracle

Quan parlem d'un oracle ens referim a un sistema que obtén el millor resum extractiu possible. El seu funcionament es basa a obtenir els fragments del text que aplicant la mètrica d'avaluació obtinguen la millor puntuació. La implementació d'aquest sistema es pot realitzar de múltiples maneres, ja que hi ha criteris a considerar que modificarien en gran manera el resultat. Podem posar com a exemple decidir aplicar la mètrica sobre les diferents frases del resum o sobre el resum en conjunt o si en lloc de frases completes del text original es designa altre tipus de segments per a aplicar les mètriques.

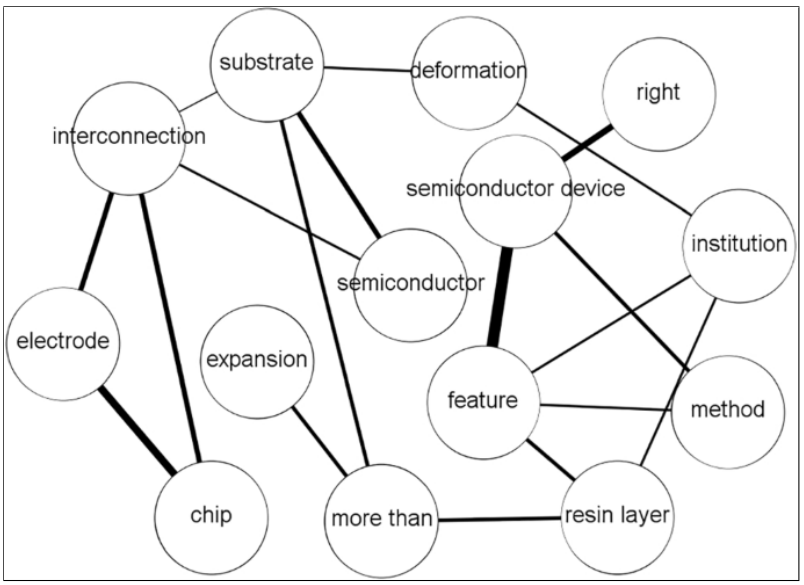

Figura 3.6: Representació del funcionament de TextRank a escala de paraula. El diferent grossor entre les arestes del graf indica el pes de la connexió entre els vèrtex.

En el present treball hem realitzat el disseny i implementació d'un oracle per a la generació de resums de notícies de manera no supervisada. L'oracle s'ha utilitzat per a maximitzar el resultat de les dues mètriques que s'utilitzen en el projecte de manera individual. La primera consideració a tindre en compte era si treballar directament amb les frases senceres del resum o utilitzar fragments de menor talla per a calcular l'oracle. Com que els sistemes extractius a utilitzar, a excepció de TextRank, es centraven en l'extracció de frases senceres que l'oracle desenvolupat treballe també d'aquesta manera sembla més que suficient per a poder garantir que s'establirà una cota superior. En segon lloc s'ha de decidir com realitzar les comparacions entre el resum de referència i el resum de l'oracle que s'ha de generar. Arran d'aquesta consideració es van presentar tres aproximacions diferents.

La primera d'elles consisteix a seleccionar el conjunt de frases del document que, en aplicar les mètriques, obtinguen la similitud més gran de manera progressiva fins que en augmentar el nombre de frases no millore la puntuació. Aquest oracle haguera assegurat una molt bona aproximació a la màxima puntuació que es pot obtenir per a cada una de les mesures mitjançant l'extracció literal de frases del document. El principal problema d'aquesta solució és la clara exponencialitat del seu cost temporal, ja que podia aplegar a ser de $\Theta(n^n)$, sent n el nombre de frases d'una notícia. A més, per una execució eficient quant a recorregut de les frases s'hauria d'emmagatzemar totes les combinacions que s'han generat en la iteració anterior (en la que s'utilitzava una frase menys).

Tenint açò en compte el cost espacial seria també de l'ordre $\Theta(n^n)$ per tal de no empitjorar el cost temporal. Com podem veure aquestes mesures són inabastables per a cap equip disponible i menys considerant el nombre de notícies amb el que treballem. A més a més, les mesures en què treballem tenen un cost de càlcul relativament elevat, principalment BERTScore, per la qual cosa aplicar tantes vegades aquesta mesura elevaria el temps d'execució a uns valors inviables. És per aquests motius que aquesta precisa, però costosa aproximació no ha sigut implementada.

La segona aproximació consistia a realitzar un oracle que seguira l'anterior esquema però limitant el nombre de frases a incloure en el resum generat al nombre de frases del resum. Aquesta solució pareixia ser més raonable a causa del reduït nombre mitjà de frases que presenten els resums dels articles utilitzats. Tot i aquesta reducció hi ha casos

que el mateix resum de la notícia està format per un molt elevat nombre de frases fent la solució de nou inviable.

En tercer lloc tenim l'alternativa per la qual finalment s'ha decidit optar. Aquesta consisteix a ajuntar les diferents frases que obtenen el millor valor en cada mesura calculant-les amb cadascuna de les frases del resum per separat. D'aquesta manera per a cada notícia el nombre de vegades que s'ha d'aplicar cada mètrica és de $n * m$, sent n el nombre de frases del document i m el nombre de frases de la notícia. Per tant, la descripció formal de l'oracle que hem implementat i fet servir és la següent:

$$Resum\ generat = \cup_{r_i \in resum} Max(Metrica(r_i, s) : s \in noticia)$$

3.5 Models basats en xarxes neuronals

A continuació passem a presentar dues de les arquitectures destacades que hi trobem en l'actualitat en la investigació sobre el tractament de textos amb bons resultats i gastant tecnologies *state-of-the-art* que hem anat comentant al llarg de l'anterior capítol i que utilitzem en aquest treball: mBART i SHANN.

3.5.1. mBART

BART [27] és una arquitectura que es basa en els *transformers sequence-to-sequence*[3] estàndard amb codificació bidireccional sobre documents. Aquests documents són modificats mitjançant diferents tècniques, com la supressió d'elements, emmascarament de paraules, permutació de paraules i oracions i rotació de documents[4], amb un codificador son completats o corregits per un sistema de descodificació per a entrenar el sistema amb la finalitat d'acomplir diferents tasques de PLN com poden ser la classificació de paraules, la traducció de textos o generació de frases.

BART és particularment efectius en tasques relacionades amb la generació de text, però també presenta bons resultats en la comprensió d'aquest. Per a tasques com diàlegs abstractius, cerca de respostes i generacions presenta resultats que serveixen de referent i comparables amb RoBERTa [42], una optimització del sistema BERT explicat anteriorment.

MBART [41] és un *auto-encoder* creat a partir desenvolupat partint de BART preentrenat amb un conjunt de corpus monolingües per a l'eliminació d'interferències en textos. Inicialment es va pensar per a treballar en tasques de traducció. En aquest treball es pretén fer servir models entrenats amb mBART per tal d'obtenir resums abstractius (ja que són fragments de text recompostos pel model i no extrets directament del document original) de les notícies de DACSA.

3.5.2. SHANN

SAHNN [19], o *Siamese Hierarchical Neural Networks* és un sistema per a la generació automàtica de resums abstractius de documents. Aquest sistema utilitza *Hierarchical Attention Networks* per a determinar, ja que l'atenció és a escala de frase, quines són les frases més importants d'un text i generar el resum amb elles.

[3] Podeu consultar l'article [11] per saber més d'aquest sistema.

[4] Es selecciona una paraula del document com el nou inici del text i es reordena el document afegint al final el text anterior a l'inici de la selecció.

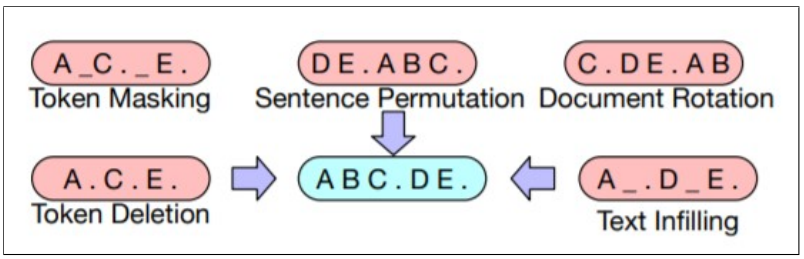

Figura 3.7: Exemple del funcionament de la generació de soroll per part del sistema de codificació de BART [27].

De manera simplificada, el que aprén aquest sistema al ser entrenat és si un resum és correcte per a un document o no. Determina que és adequat quan troba certa similitud semàntica. L'entrenament consisteix a entrenar xarxes neuronals siameses mitjançant parells de document-resum per tal de determinar si un resum és adequat per a un document donat. A aquest entrenament se li suma un mecanisme d'atenció basat en *Hierarchical Attention Networks* per tal d'assignar un pes a les frases del document, la qual cosa permet ordenar-les i generar un resum amb les n primeres frases seleccionades.

Aquest entrenament requereix representacions de paraules mitjançant vectors per tal d'establir la representació de documents en el mecanisme d'atenció. En aquest treball s'utilitzaran *embeddings* entrenats a partir del mateix corpus DACSA i documents extrets de la Viquipèdia.

CAPÍTOL 4
Ferramentes utilitzades

Com ja hem esmentat en la memòria, hi ha ferramentes que faciliten en gran manera la implementació i el desenvolupament del treball. En aquest capítol descrivim els recursos software i hardware que necessaris per al desenvolupament de l'experimentació, explicant el seu funcionament i quines funcionalitats seran d'utilitat al llarg del projecte.

4.1 Software

A continuació expliquem de quina manera són rellevants els recursos software que esmentem per a la realització d'aquest treball i de quina manera han sigut utilitzats en la realització de les tasques.

4.1.1. Python

Python en un llenguatge de programació interpretat, d'alt nivell, orientat a objectes i de propòsit general. Donat que un dels seus objectius principals és garantir la llegibilitat del codi resulta molt senill d'utilitzar per a iniciar-se en la programació. Ja que no s'especialitza en una única tasca no sol ser el millor llenguatge quant a eficiència, però disposa d'una molt gran varietat de ferramentes que permeten desenvolupar codi per a la majoria de tasques habituals presents en la informàtica a alt nivell.

El motiu pel qual decidim fer ús d'aquest llenguatge és que disposa d'una gran quantitat de llibreries útils per al tractament de textos i de xarxes neuronals. Per aquest motiu també Python és actualment un dels llenguatges de programació més utilitzats en àrees de recerca en aprenentatge automàtic. Un gran percentatge dels treballs desenvolupats presenten els seus resultats o implementació en formats compatibles amb Python. També les bases de dades més comunes presenten estructures compatibles o úniques de Python.

En el nostre cas hem fet servir des de llibreries bàsiques de Python com NumPy, CSV, jsonlines o statistics per tal de realitzar operacions secundàries de manera més ràpida. Un sistema que hem fet servir també mitjançant el llenguatge de Python ha sigut el model de SHANN. Les llibreries més avançades de les que hem fet ús són les que comentarem a continuació les expliquem a continuació en aquest mateix capítol.

4.1.2. NLTK

Natural Language Toolkit és una ferramenta destinada al tractament de documents en llenguatge natural en Python. Aquesta eina disposa de funcionalitats per a classificació, tokenitzat, derivació de paraules (*stemming*) i etiquetat entre altres. En aquest treball hem uti-

litzat la funcionalitat de word_tokenizer de la llibreria NLTK (junt amb altres de manera auxiliar com emoticon) per tal de tokenitzar les notícies i els resums de referència per paraules per a l'entrenament tant dels *embeddings* com de SHANN.

4.1.3. SpaCy

La llibreria de SpaCy, sota llicència del MIT, té la finalitat de proporcionar una ampla gamma de ferramentes destinades al PLN avançat. D'igual manera que NLTK està creada per al seu ús en Python principalment. El motiu pel qual s'ha utilitzat en aquest treball és per poder obtenir models de representació de paraules necessàries per a la utilització d'altres ferramentes com PyTextRank.

4.1.4. TensorFlow

TensorFlow és conegut per la gran varietat d'utilitats que presenta per al desenvolupament de ferramentes d'aprenentatge automàtic. Es tracta d'una llibreria de Python de codi obert destinada a accelerar processos de computació numèrica. Aquesta permet crear ràpidament models d'aprenentatge profund i entrenar-los i utilitzar-los amb gran facilitat. La utilització de TensorFlow en aquest treball la trobem completament en la configuració, entrenament i testatge del model de SHANN.

4.1.5. Gensim

Gensim és una llibreria per a Python que realitza l'entrenament no supervisat de representacions de dades a partir de documents de text de manera eficient. Gensim s'ha utilitzat per a entrenar els *embeddings* utilitzats en l'entrenament de SHANN.

Aquest entrenament s'ha realitzat utilitzant Word2Vec, tant per a català com per a castellà. Els documents a partir dels quals han sigut entrenats són la Viquipèdia i els conjunts d'entrenament de DACSA tokenitzats amb l'anteriorment esmentada funcionalitat de NLTK. Tots els models són del tipus skip-grama amb vectors de dimensió 300. Les principals diferències entre els *embeddings* recau en el nombre d'aparicions que requereix una paraula er a ser inclosa en el model (1, 10, 100...) i si s'ha eliminat o no la presència de majúscules en el corpus abans de ser entrants.

4.2 Hardware

Tot i que en el present treball no es fa cap estudi ni implementació relacionada amb el sector del hardware cal destacar la importància de poder fer ús de targetes gràfiques per a la realització de les diferents tasques del projecte. Les xarxes neuronals, de la mateixa manera que altres ferramentes de les quals fem ús en aquest projecte, depenen per a la seua utilització o entrenament d'un gran nombre de menudes operacions bàsiques independents. Aquestes operacions es poden optimitzar per tal de ser operades en forma matricial i accelerar el procés.

Per aquest motiu les targetes gràfiques tenen una gran rellevància a l'hora de realitzar recerques relacionades amb xarxes neuronals. Gràcies al fet que operen amb estructures matricials i estan destinades a realitzar aquest tipus d'operació són capaces d'accelerar els càlculs necessaris en molts processos en que es veuen involucrades les xarxes neuronals.

En el nostre cas hem fet ús de Cuda (Cuda 11), una ferramenta de NVIDIA que permet utilitzar les targetes gràfiques d'aquesta marca per a la realització d'operacions matricials

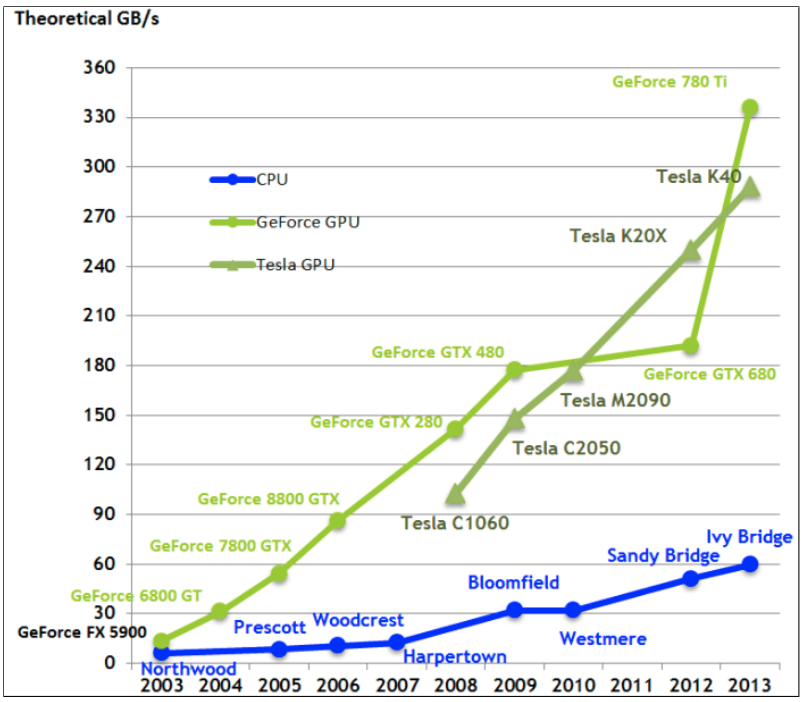

Figura 4.1: Comparativa de velocitat entre GPU i CPU on es veu clarament l'acceleració que pot aportar l'ús de targetes gràfiques en treballar amb xarxes neuronals.

d'una manera senzilla. Les targetes gràfiques que s'han utilitzat en aquest projecte han sigut una NVIDIA Titan X, proporcionada per la màquina BOSO del grup d'investigació i una NVIDIA GEFORCE GTX 1650.

CAPÍTOL 5
Experimentació i resultats

En el present capítol descrivim l'experimentació que s'ha dut a terme durant la realització d'aquest treball. Començarem definint els diferents aspectes del corpus DACSA que s'han avaluat per tal de prendre decisions en l'experimentació. Seguidament descriurem el procés d'obtenció de resums, tractarem els mètodes tradicionals no supervisats seguit dels resums extractius obtinguts mitjançant SHANN i els abstractius generats amb mBART. Finalment, compararem els resultats i valorarem el procés que s'ha realitzat durant l'experimentació.

5.1 Anàlisi del corpus

Com hem comentat amb anterioritat és necessari disposar d'un corpus de qualitat per tal d'assegurar que l'avaluació dels resultats obtinguts és vàlida. En primer lloc, ja que fem ús de la mesura lead-2 per a la generació de resums calia conéixer si la partició, i més concretament, les diferents fonts de DACSA tenien una puntuació excessivament elevada amb aquesta mesura.

El motiu pel qual és necessari fer aquesta comprovació és que, com es comenta al capítol 3, en molts articles periodístics presenten al començament del cos la informació més important de la notícia i que fins i tot, en alguns casos, s'utilitza aquest començament com resum associat a la notícia.

Per a fer aquesta comprovació s'ha obtingut els distints valors de ROUGE (multiplicats per 100 per poder realitzar una millor comparació) sobre els resums de referència de les notícies i les dues primeres frases d'aquestes. Sobre els resultats, els quals trobem a les taules 5.1, 5.2, 5.3 i 5.4 podem realitzar diverses afirmacions.

En primer lloc cal comentar que el resultat mitjà de les fonts és molt similar independentment de la partició que s'estudie. Aquest fet ens dóna una idea que la separació de les particions és consistent i no es troba cap anomalia notòria en elles.

Pel que fa al conjunt en català veiem que totes les fonts presenten unes puntuacions similars les quals es poden considerar acceptables. El fet que totes presenten en ROUGE-1 un valor proper a 20 ens fa saber que, com és natural en una notícia, en les primeres frases s'ha parlat sobre el que tracta la notícia (segons el seu resum) però no representa la totalitat del resum de referència. L'única font amb un valor de ROUGE-1 més elevat és *VilaWeb* amb una puntuació de quasi 45. Tot i això, com es tracta d'una font que només es troba a la partició de *hard test* no tindrà cap impacte a l'entrenament de models.

Per altra banda, en el conjunt de castellà observem alguns valors superiors als que presentava el conjunt en català. Tot i això, l'única font amb un valor destacable dins de

les particions de *train*, *test* i *valitation* és *El Mundo*. A la taula 3.2 veiem que *El Mundo* representa 35 112 de les 2 136 910 notícies que finalment formen part del conjunt de notícies en castellà. Com que representa menys del 2% considerem que no suposarà un impacte rellevant a l'hora de generar els resums. En la partició de *hard test* sí que observem vàries fonts amb valors de ROUGE elevats com *Agencia EFE* o *El Economísta* que d'igual manera que *Vilaweb* en el conjunt en català, no pot suposar cap inconvenient a la generació de resums de les particions principals ni a l'entrenament de models.

Font Periodistica	**Test**				**Hard Test**			
	R1	R2	RL	RLS	R1	R2	RL	RLS
Diari ARA	22,20	08,80	16,82	18,44				
Diari de Griona	25,14	10,35	18,55	20,32				
Diari La Veu	22,07	09,96	16,82	17,84				
El Per. de Catalunya	24,01	08,99	17,16	19,66				
El Punt Avui					27,04	10,77	19,14	21,09
El Temps					25,87	07,00	15,91	20,48
Nació Digital	22,26	07,16	15,80	17,53				
Regió 7	24,83	10,69	18,53	20,45				
VilaWeb					44,63	33,61	39,56	41,09

Taula 5.1: Resultats de l'anàlisi de puntuació lead-2 per fonts per a les particions de *test* i *hard test* en català.

Font Periodistica	**Train**				**Validation**			
	R1	R2	RL	RLS	R1	R2	RL	RLS
Diari ARA	22,06	08,79	16,80	18,36	22,01	08,75	16,78	18,28
Diari de Griona	25,32	10,45	18,60	20,45	25,44	10,62	18,71	20,58
Diari La Veu	22,03	09,91	16,75	17,81	22,14	10,13	16,85	17,87
El Per. de Catalunya	23,80	08,79	16,97	19,41	23,82	08,82	17,06	19,47
El Punt Avui								
El Temps								
Nació Digital	22,41	07,14	15,78	17,55	22,35	07,13	15,78	17,57
Regió 7	24,67	10,58	18,55	20,47	24,58	10,43	18,46	20,34
VilaWeb								

Taula 5.2: Resultats de l'anàlisi de puntuació lead-2 per fonts per a les particions de *train* i *validation* en català.

Altra dada que s'ha volgut obtindre sobre aquest corpus era el cobriment del vocabulari que presenten els diferents *embeddings* disponibles per a l'entrenament de models. Per a realitzar aquesta comprovació s'ha recorregut tot el corpus de noticies i s'ha comparat quantes paraules de cada partició estaven presents en el conjunt d'*embeddings*. Aquest valor és important ja que ajudarà a entendre millor els resultats que es generen en els models que fan ús dels *embeddings*. En cas de presentar un baix cobriment seria més probable que els resultats no foren massa prometedors, ja que gran part del vocabulari seria nou i els mecanismes d'atenció no funcionarien adequadament.

Els cobriments dels diferents conjunts d'*embeddings* que s'han utilitzat els podem trobar a les taules 5.5, 5.6, 5.7 i 5.8. Tant en català com en castellà observem un cobriment del vocabulari de més del 99% en tots els conjunts d'*embeddings*.

Font Periodistica	Test				Hard Test			
	R1	R2	RL	RLS	R1	R2	RL	RLS
20 Minutos					33,25	08,96	19,62	22,77
ABC	25,93	11,52	19,65	21,39				
Diario de Mallorca	25,32	11,64	19,58	21,14				
Agencia EFE					59,88	52,54	57,24	58,19
El Diario	34,06	13,55	22,14	25,20				
El Economísta					68,59	65,28	66,85	67,98
El Español	26,81	09,61	18,75	21,70				
El Independiente					26,62	10,22	19,23	21,04
El Mundo	61,14	52,30	58,07	58,86				
El País	26,31	11,98	20,35	22,35				
El Per. de Aragón	24,68	10,93	18,94	20,87				
El Per. de Catalunya	24,57	08,51	17,45	18,89				
Europa Press					22,70	14.77	20,18	20,72
Expansión					22,94	06,34	17,26	19,77
La Razón	25,28	11,88	20,27	20,97				
Las Provincias	22,91	09,96	18,17	19,21				
La Vanguardia	24,02	09,12	17,77	19,93				
Levante	25,44	11,30	19,42	20,98				
Nodo 50					25,33	06,19	16,00	20,37
Público	30,57	14,71	22,72	25,37				
Última Hora					33,80	20,67	27,93	29,74

Taula 5.3: Resultats de l'anàlisi de puntuació lead-2 per fonts per a les particions de *test* i *hard test* en castellà.

Font Periodistica	Train				Validation			
	R1	R2	RL	RLS	R1	R2	RL	RLS
20 Minutos								
ABC	25,80	11,28	19,49	21,23	25,73	11,12	19,31	21,08
Diario de Mallorca	25,32	11,63	19,53	21,15	25,45	11,64	19,56	21,18
Agencia EFE								
El Diario	33,97	13,60	22,15	25,18	33,89	13,48	22,10	25,10
El Economísta								
El Español	26,81	09,49	18,69	21,64	26,72	09,48	18,67	21,55
El Independiente								
El Mundo	60,64	51,73	57,47	58,29	60,78	61,93	57,69	58,50
El País	26,12	11.87	20,20	22,19	26,12	11,75	20,15	22,09
El Per. de Aragón	24,90	11,11	19,10	21,08	24,80	10,92	18,94	21,18
El Per. de Catalunya	24,74	08,61	17,60	20,07	24,72	08,65	17,67	20,13
Europa Press								
Expansión								
La Razón	25,60	12,28	20,67	21,33	25,46	12,23	20,59	21,24
Las Provincias	22,89	09.95	18,16	19,19	22,88	10,10	18,19	19,24
La Vanguardia	23,84	08,90	17,60	19,72	23,77	08.85	17,57	19,70
Levante	25,36	11,16	19,28	20,86	25,36	11,06	19,27	20,85
Nodo 50								
Público	30,35	14,35	22,41	25,04	30,16	14,14	22,17	24,85
Última Hora								

Taula 5.4: Resultats de l'anàlisi de puntuació lead-2 per fonts per a les particions de *train* i *validation* en castellà.

La diferència més destacable entre els *embeddings* que només han generat vectors de paraules amb un mínim de 10 aparicions i els que només requerien 1[1] és que, encara que el cobriment de vocabulari és molt pitjor, el cobriment de tokens[2] aparegut en les diferents particions és mantén molt similar. S'entén amb aquestes dades que un elevat percentatge del vocabulari apareix en molt poques ocasions. El poc pes que presenta un nombre tan elevat del vocabulari presenta la possibilitat que l'entrenament es veja entorpit per aquestes paraules minoritàries. Tot i això, l'alt percentatge de cobriment de token indica que els *embeddings* sí que són adequats per a l'entrenament de models.

Partició	Tokens	Vocabulari
Test	99,77%	89,05%
Hard Test	99,74%	86,43%
Train	99,93%	97,27%
Validation	99,78%	89,43%

Taula 5.5: Percentatge de cobriment de token i de vocabulari que presenten els *embeddings* en català, sense eliminar majúscules i amb 1 aparició per paraula per a incloure als *embeddings*.

Partició	Tokens	Vocabulari
Test	99,36%	69,92%
Hard Test	99,33%	68,92%
Train	99,43%	30,72%
Validation	99,39%	70,56%

Taula 5.6: Percentatge de cobriment de token i de vocabulari que presenten els *embeddings* en català, sense eliminar majúscules i amb 10 aparicions per paraula per a incloure als *embeddings*.

Partició	Tokens	Vocabulari
Test	99,86%	87,95%
Hard Test	99,85%	87,23%
Train	99,95%	96,80%
Validation	99,86%	87,96%

Taula 5.7: Percentatge de cobriment de token i de vocabulari que presenten els *embeddings* en castellà, sense eliminar majúscules i amb 1 aparició per paraula per a incloure als embeddings.

Partició	Tokens	Vocabulari
Test	99,63%	69,19%
Hard Test	99,64%	70,49%
Train	99,60%	29,28%
Validation	99,64%	69,23%

Taula 5.8: Percentatge de cobriment de token i de vocabulari que presenten els *embeddings* en castellà, sense eliminar majúscules i amb 10 aparicions per paraula per a incloure als *embeddings*.

Finalment sobre el corpus s'ha obtingut el nombre mitjà de tokens i frases per notícies i resums de referència, ja que era un paràmetre important per a l'entrenament del model de SHANN. Observem en els valors presentats a la taula 5.9 que els valors mitjans de longitud de frases i paraules dels parells de notícies i resums són similars per a ambdós idiomes.

Llengua	Frases Notícies	Tokens Notícies	Frases Resums	Tokens Resums
Català	19,17	24,36	1,23	18,51
Castellà	25,43	22,43	1,32	19,47

Taula 5.9: Valors de les mitjanes de llargària en paraules i frases per a les notícies i resums del conjunt de dades en català i castellà.

[1] Notem que, tot i que els *embeddings* han sigut entrenats amb el conjunt de *train* i només requerint una aparició d'una paraula per a tractar-la als *embeddings* no presenten un 100% de cobriment del vocabulari d'aquest conjunt. Donat que la tokenització del corpus en generar els conjunts d'*embeddings* s'ha realitzat en un moment i equip diferent de la tokenització realitzada en realitzar els càlculs de cobriment s'explica l'existència de vocabulari no cobert per la diferència de versions de les ferramentes utilitzades.

[2] Amb token fem referència a cada un dels elements que componen les frases d'un document. El diferenciem del concepte de *paraula* amb el qual ens referim a cada una dels elements que componen el vocabulari o, dit d'altra manera, cada un dels tokens únics presents a un document.

5.2 Obtenció de resums

D'acord amb els resultats de l'anterior anàlisi s'han determinat alguns dels paràmetres a utilitzar en la generació de resums d'alguns sistemes. A continuació passem a explicar els processos d'obtenció de resums realitzats i presentar els resultats de cadascun d'ells. Els resums i resultats en cada sistema s'han obtingut tant per al conjunt de *test* com per al conjunt de *hard test*.

5.2.1. Sistemes tradicionals no supervisats

Com s'ha introduït en capítols anteriors de la memòria els sistemes tradicionals que s'han utilitzat en aquest treball són un Oracle (desenvolupat i implementat en el mateix treball), TextRank i Lead-k.

Oracle

L'oracle utilitzat és exactament el descrit a la metodologia del treball. S'ha obtingut una frase del document per cada una de les frases del resum. Per a seleccionar aquestes frases és necessari establir el criteri mitjançant el qual determinar que una frase és millor que una altra. Com en el present treball hem fet servir dues mètriques per a l'obtenció de resultats, ROUGE i BERTScore, s'ha decidit implementar dos oracles diferents, un pensat per obtenir els millors[3] resums mitjançant ROUGE-Lsum i l'altre per obtenir els millors resums segons BERTScore. Com a resultat tenim les taules 5.10 i 5.11 per a l'oracle basat en ROUGE-Lsum i les taules 5.12 i 5.13 per a l'oracle basat en BERTScore.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	41,68	25,53	36,29	36,64	75,87
Hard Test	47,16	29,44	40,23	41,82	75,86

Taula 5.10: Resultats de l'oracle de ROUGE per al conjunt de notícies en català.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	46,04	30,12	40,85	41,37	77,45
Hard Test	45,49	25,50	36,84	37,54	74,85

Taula 5.11: Resultats de l'oracle de ROUGE per al conjunt de notícies en castellà.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	40,88	25,11	34,55	34,95	76,64
Hard Test	46,10	28,71	37,98	39,73	76,79

Taula 5.12: Resultats de l'oracle de BERTScore per al conjunt de notícies en català.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	45,43	30,22	39,48	40,01	78,60
Hard Test	45,18	25,88	35,80	36,39	76,16

Taula 5.13: Resultats de l'oracle de BERTScore per al conjunt de notícies en castellà.

[3]Remarquem que la definició de millor en aquest context s'adapta la definició de l'oracle implementat, no al millor resum real.

TextRank

Per a l'ús de TextRank s'ha fet servir l'implementació proporcionada en la llibreria PyTextRank. Aquesta ha sigut utilitzada sense cap paràmetre adicional ja que a l'hora de comparar els comportaments s'ha observat que les puntuacions son majors d'aquesta manera.[4]

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Test	16,54	5,62	11,98	15,33
Hard Test	17,16	5,83	12,27	15,93

Taula 5.14: Resultats de TextRank per al conjunt de notícies en català.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Test	14,13	4,27	8,13	13,15
Hard Test	21,78	6,13	11,77	18,97

Taula 5.15: Resultats de TextRank per al conjunt de notícies en castellà.

Lead-K

Finalment, dels sistemes no supervisats, s'ha obtingut els resums mitjançant Lead-K. Tot i que la mesura més popular és probablement Lead-3, donat que la mitjana de la longitud en frases dels resums no aplega a 2 tant en català com en castellà (taula 5.9), s'ha decidit que la millor opció per a realitzar les comparacion seria utilitzar la mesura de Lead-2. A les taules 5.16 i 5.17 es poden trobar els resultats de les mètriques aplicades sobre els resums generats per Lead-2.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	23,41	9,33	17,28	19,04	68,96
Hard Test	31,44	15,74	23,63	26,32	70,30

Taula 5.16: Resultats de Lead-2 per al conjunt de notícies en català.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	29,00	14,39	22,56	24,45	71,03
Hard Test	36,64	16,79	26,07	28,64	71,81

Taula 5.17: Resultats de Lead-2 per al conjunt de notícies en castellà.

5.2.2. Sistemes basats en xarxes neuronals

Una vegada obtinguts els resums dels models tradicional passem a descriure l'obtenció de resums a partir de models basats en xarxes neuronals. En aquesta secció presentem tant extractius, generats a partir de SHANN, com abstractius proporcionats per mBART.

SHANN

En l'obtenció de resums a partir de SHANN ha sigut necessari determinar una sèrie de paràmetres bàsics per a l'obtenció dels resums. D'acord amb les mitjanes de llargària ob-

[4] Els valors que falten no han pogut ser obtinguts abans de la redacció de la memòria a causa dels recursos i el temps necessaris per a la seua obtenció.

tingudes en l'estudi de corpus DACSA s'ha establert els valors màxims que havia d'esperar el sistema a l'hora de l'entrenament. Els valors dels paràmetres ha quedat de la següent manera:

- **max_len_doc_sents** : 25
- **max_len_doc_sent_words** : 30
- **max_len_summ_sents** : 2
- **max_len_summ_sent_words** : 30.
- **topk_sentences** : 2.

A més a més, SAHNN requereix d'*embeddings* tant per a l'entrenament com posteriorment per a la generació dels resums. En el nostre cas hem disposat per a cada idioma de fins a quatre conjunts d'*embeddings*, primerament variant entre la presència o absència de majúscules i i en segon lloc depenent de si es treballa amb 1 o 10 com a nombre mínim d'aparicions d'una paraula per a ser inclosa al vocabulari dels conjunts d'*embeddings*. A causa dels prolongats temps d'entrenament dels models només s'han utilitzat els *embeddings* en els quals sí que es distingeix entre majúscules i minúscules, resultant en dos models d'*embeddings* per a cada llengua.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	23,03	9,28	16,86	18,67	68,54
Hard Test	29,68	12,82	21,13	24,12	69,35

Taula 5.18: Resultats de SHANN per al conjunt de notícies en català utilitzant *embeddings* amb distinció de majúscules i un mínim d'aparicions de 1.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	26,71	11,71	19,89	21,82	69,77
Hard Test	35,67	15,41	24,76	27,53	70,90

Taula 5.19: Resultats de SHANN per al conjunt de notícies en castellà utilitzant *embeddings* amb distinció de majúscules i un mínim d'aparicions de 1.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	23,56	9,64	17,31	19,16	68,85
Hard Test	30,40	13,79	22,01	24,93	69,72

Taula 5.20: Resultats de SHANN per al conjunt de notícies en català utilitzant *embeddings* amb distinció de majúscules i un mínim d'aparicions de 10.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	26,93	11,84	20,07	22,12	69,86
Hard Test	35,55	15,22	24,63	27,41	70,83

Taula 5.21: Resultats de SHANN per al conjunt de notícies en castellà utilitzant *embeddings* amb distinció de majúscules i un mínim d'aparicions de 10.

En conseqüència, obtenim les següents taules de resultats, on les taules 5.18 i 5.19 mostren els resultats dels *embeddings* que requereixen una única aparició per paraula i les taules 5.20 i 5.21 mostren els resultats dels *embeddings* que requereixen 10 aparicions per paraula per a incloure-les al vocabulari.

mBART

El sistema que hem utilitzat per a l'entrenament de models per a la generació automàtica de resums abstractius és mBART. Com hem explicat anteriorment aquest sistema intenta recompondre text quan es produeix soroll. Partint d'aquest funcionament és capaç de generar resums sense haver d'obtindre tot el contingut del text original.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	28,58	11,89	22,99	23,39	72,02
Hard Test	27,46	11,04	21,13	22,01	70,33

Taula 5.22: Resultats de mBART per al conjunt de notícies en català.

Particio	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
Test	31,09	13,56	24,67	25,47	72,25
Hard Test	30,66	12,08	23,14	23,89	71,07

Taula 5.23: Resultats de mBART per al conjunt de notícies en castellà.

5.3 Valoració de l'experimentació

A continuació interpretarem i compararem els resultats de l'avaluació dels resums dels diferents mètodes que s'han utilitzat en l'experimentació. En primer lloc tractarem sobre els resums generats a partir de sistemes tradicionals no supervisats i seguidament valorarem els resultats obtinguts a partir de sistemes basats en xarxes neuronals.

Quant als resultats obtinguts dels sistemes tradicionals podem destacar la proximitat que presenten els valors de les mètriques d'ambdues implementacions de l'oracle. Encara que en l'oracle de ROUGE no s'ha utilitzat la mesura de BERTScore per a l'obtenció dels resums presenten un valor lleugerament inferior en aquesta mesura a l'obtingut per l'oracle de BERTScore. El mateix ocorre a l'oracle de BERTScore però canviant la mètrica. Podem deduir amb aquestes dades que les dues mesures estan interrelacionades o, almenys, que a major coincidència sintàctica es tendeix a obtindre una major similitud semàntica i viceversa.

Per altra banda, s'observa com tant en català com en castellà els valors de Lead-2 són molt més elevats en el conjunt de *hard test* que en el de *test*. Amb aquesta informació es reafirma la importància de distingir aquestes dues particions, ja que són valors propers als obtinguts per l'oracle i que podrien estar en una situació avantatjada a l'hora de comparar amb altres sistemes de generació automàtica de resums.

Finalment, s'observa també una gran diferència entre els resums obtinguts per TextRank i els obtinguts a partir de Lead-2. El fet que es tracte d'articles periodístics justifica aquesta gran diferència, ja que no deixen de presentar informació important de la notícia en les primeres línies del document. Front a la gran separació de resultats en sistemes tradicionals podem establir TextRank com una cota inferior a superar, Lead-2 com una referència amb bons resultats a superar i els resultats de l'oracle com la cota superior, ja que aplegar al nivell dels oracles voldria dir que el sistema extractiu és quasi ideal sense tindre accés als resums de referència.

A continuació passem a comentar els resultats obtinguts a partir de sistemes basats en xarxes neuronals. Primerament trobem els resums obtinguts mitjançant l'ús de SHANN. En primer lloc observem que un major recobriment del vocabulari en els *embeddings* no implica necessàriament millors resultats. Tot i que es tracta d'una lleugera diferència ens

permet saber que és necessari valorar l'ús d'*embeddings* entrenats de diferents maneres per tal de determinar quins són els millors paràmetres.

Prenent de referència els resultats obtinguts usant els *embeddings* que necessiten un mínim de 10 aparicions, en el conjunt de *test* podem veure com amb la mesura de ROUGE s'aplega a superar els resultats obtinguts per Lead-2. Per a la resta de valors d'aquest conjunt veiem que els valors són inferiors, però no presenten una gran diferència. També per al conjunt de *hard test* observem que, encara que segueixen sent valors inferiors als obtinguts pels resums generats amb Lead-2 trobem que segueixen estant per valors propers a aquests. S'ha de tindre en consideració que es tracta de fonts no vistes en l'entrenament i que presenten valors molt més elevats a la mètrica de ROUGE respecte al conjunt de *test* en els resultats que fan ús del sistema Lead-2.

Quant a la similitud semàntica observem que en tots els casos s'obtenen pitjors valors que amb els resums de Lead-2. Tot i això observem que es tracta de valors molt semblants en els corresponents conjunts. Es pot pensar que el motiu d'aquest comportament es pot deure a canvis a la tokenització a l'hora de reescriure els resums per part dels models de SHANN.

Finalment obtenim els resultats aportats pels resums de mBART. En els valors obtinguts observem que per a les particions de *train* els resultats són millors en els resums generats amb models mBART que els obtinguts per Lead-2 tant en català com en castellà. L'única puntuació que és lleugerament inferior en els resums de mBART que en els resums de Lead-2 és ROUGE-2 en el cas del castellà. Com hem comentat al llarg de la memòria aquesta puntuació està molt relacionada amb la llegibilitat d'un text. El fet que resums abstractius obtinguen una puntuació tan similar o superior a fragments extrets del mateix document proporciona una idea de la qualitat del resum generat.

Una altra dada destacable és que mentre que en castellà els valors són lleugerament millors als obtinguts per Lead-2 en català la diferència és molt major, superant en més de 4 punts la mesura de ROUGE-Lsum, en prop de 5 punts les mesures de ROUGE-1 i ROUGE-L i en quasi 3 punts la mesura de BERTScore. Amb aquestes dades s'extrau que es tracten de resums generalment millors que els extrets mitjançant Lead-2.

Quant a la partició de *hard test* observem que, tot i ser un conjunt de dades on Lead-2 es veu afavorit i un conjunt de fonts que no s'ha vist durant l'entrenament del model, les puntuacions no són molt inferiors per a la mètrica de ROUGE i són comparables a les presentades per Lead-2 per al conjunt de *test*. Altrament, les puntuacions de BERTScore indiquen que la similitud semàntica d'ambdues generacions està prou igualada quan es compara amb els resums de referència.

Per tal de poder il·lustrar el comportament dels sistemes de generació automàtica de resums utilitzats, exposem a l'apèndix A. per a cada llengua, un exemple de notícia i els resums generats pels diferents sistemes.

En aquesta valoració hem vist que els resums obtinguts a partir de sistemes basats en xarxes neuronals tenen resultats iguals o millors als obtinguts per alguns sistemes tradicionals no supervisats. Amb açò es referma que la utilització d'aquests sistemes pot aplegar a obtenir resums de qualitat elevada. A més a més, el sistema que millors resultats a obtingut, excloent els oracles, ha sigut mBART, el qual és l'únic que genera resums abstractius. Finalment, podem concloure que DACSA podria considerar-se un corpus de qualitat d'acord amb els resultats obtinguts amb un preprocessat mínim de les dades.

CAPÍTOL 6
Conclusions

Al llarg d'aquesta memòria hem vist les diferents decicions preses al llarg del treball, el funcionament dels sistemes utilitzats per a la generació de resums i les puntuacions mitjanes obtingudes pels resums en les dues mètriques exposades. Hem vist la descripció del corpus de dades que s'ha utilitzat per a la realització i avaluació de resums, DACSA, i un estudi de diferents característiques d'aquest corpus. Com que, a més, els valors obtinguts dels estudis han sigut utilitzats a la resta del projecte per tal de prendre decisions podem considerar que es compleix el primer objectiu del treball, **Anàlisi de les característiques pròpies del corpus**. El segon objectiu, **Generació de resums fent ús de tècniques no supervisades**, també s'ha complit totalment, ja que hem fet servir tres tècniques tradicionals distintes, a partir d'implementació pròpia o amb ferramentes ja existents, que han servit per a obtenir resums diferents del conjunt de notícies treballat.

A banda d'aquests mètodes hem entrenat models SHANN i mBART per tal d'obtenir resums extractius i abstractius mitjançant sistemes basats en xarxes neuronals exposats al llarg de la memòria. Tot i això, encara que l'objectiu tercer, **utilitzar sistemes basats en xarxes neuronals per a la generació de resums**, s'ha aconseguit donat que hem realitzat el que proposàvem al començament del treball, per motius que explicarem de manera més esplaiada a la següent secció del present capítol, hi ha hagut altres sistemes que s'han quedat sense ser utilitzats per a la generació de resums. L'objectiu d'**avaluar els resums generats en els distints processos** s'ha complit completament, ja que de tots els resums generats s'han avaluat mitjançant diferents mètriques. A més a més, les mètriques utilitzades, ROUGE i BERTScore, compleixen a l'hora de mesurar les propietats sintàctiques i semàntiques dels resums generats en comparació als resums de referència.

A continuació presentem els principals entrebancs que han sorgit durant el desenvolupament de tot el projecte i a l'hora d'obtenir els resultats i les possibles línies de treball que es poden seguir partint del treball que acabem d'exposar.

6.1 Problemes sorgits

Durant l'obtenció de resums i l'estudi del corpus i els resums generats han aparegut una serie d'esdeveniments que han produït canvis en les tasques a realitzar o han dificultat o endarrerit la realització d'aquest treball. En aquesta secció enumerem alguns d'aquests sucesos que han influenciat en el treball i com han afectat a les diferents tasques.

- **Modificacions del corpus**: durant la realització d'aquest treball s'ha hagut de generar resums mitjançant diferents sistemes i estudiar les característiques del conjunt de notícies vàries vegades. El motiu de la repetició dels experiment es deu al fet

que el corpus DACSA va rebre diverses modificacions una vegada iniciada l'experimentació. Entre aquestes modificacions trobem canvis sobre el conjunt de notícies que finalment s'inclouen en el corpus, creació de nou de les diferents particions junt amb la decisió de crear un conjunt a banda, la partició de *hard test* amb fonts diferents, i el canvi en el format que es presentava DACSA, afectant així a la manera d'accedir a les dades i generant la necessitat d'adaptar el codi generat per a l'execució de les diferents proves.

- **La beca on s'enmarca el treball**: la línia de recerca de la beca en la qual s'ha dut a terme el treball s'ha encaminat principalment a l'estudi de DACSA i la realització de proves amb generació de resums amb ferramentes principalment extractives. Per aquest motiu no s'han realitzat tants càlculs sobre resums abstractius com s'havia pensat a l'hora de proposar el treball.

- **Temps d'execució**: les tasques acomplides durant el present treball han sigut notòriament costoses en termes de temps i espai. Donat que es tracta d'un conjunt de notícies i resums de grans dimensions l'avaluació de resums aplicant les diferents mètriques ha resultat especialment llargues. Gràcies a la disponibilitat de la màquina BOSO del grup d'investigació, molts d'aquests temps s'han pogut reduir. Tot i això, moltes de les execucions realitzades han tardat diversos dies a donar resultat o han necessitat un gran intercanvi de dades a través de la xarxa, produint alentiments a l'hora de poder realitzar l'experimentació.

  Altrament, la generació d'*embeddings* és una tasca no trivial que requereix un entrenament molt lent, obligant així a modificar alguns criteris d'entrenament (en lloc d'utilitzar el conjunt de dades de OSCAR, Viquipèdia i el corpus DACSA només s'ha utilitzat Viquipèdia i DACSA per la gran quantitat de temps que requeria l'entrenament incloent totes les dades. Els temps d'execució també han afectat a l'obtenció de resums amb SHANN a partir de diferents *embeddings*, ja que l'entrenament del model i la generació i avaluació de resums podia tardar més dos dies per a cada modificació dels *embeddings*. En conseqüència també s'ha vist endarrerit la producció de resums abstractius.

6.2 Treball futur

Com hem dit a l'inici del present capítol hi ha diferents tasques que es podrien haver realitzat en el treball que hem presentat per tal d'augmentar el nombre de resultats obtinguts, millorar els resultats o ampliar els estudis realitzats. Aquestes són algunes tasques o línies de recerca que es podrien realitzar per tal de continuar o complementar els estudis realitzats en aquest treball:

- **Utilització de nous models**: Com hem comentat en diferents ocasions no s'han utilitzat models utilitzant tants sistemes com s'haguera pensat en un primer moment. A més a més, alguns dels *embeddings* que s'han generat amb la finalitat de ser utilitzat han quedat sense ser estudiats per la necessitat de redactar la memòria del treball. Per aquest motiu, la recerca d'altres sistemes o continuar explorant la generació de resums mitjançant SHANN complementarien adequadament els resultats obtinguts en el present treball.

- **Modificacions sobre DACSA**: aquest treball s'ha realitzat directament sobre el corpus DACSA. Les diferents tasques s'han realitzat sense cap altre preprocès més que la tokenització amb NLTK o el pas de les lletres majúscules a lletres minúscules. Un

estudi dels símbols estranys presents al corpus depurant la tokenització o representació de les dades més enllà de la que aplica automàticament NLTK podria generar uns resultants diferents, ja que l'entrenament i captura de la relació semàntica per part dels models basats en xarxes neuronals es podria veure positivament afectada amb aquests canvis.

Bibliografia

[1] Adam Roberts, Colin Raffel i Noam Shazeer How Much Knowledge Can You Pack Into the Parameters of a Language Model? *arXiv preprint arXiv:2002.08910*, febrer, 2020.

[2] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy i Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, abril, 2018.

[3] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy i Samuel R. Bowman. SuperGLUE: A stickier benchmark for generalpurpose language understanding systems. *arXiv preprint arXiv:1905.00537*, maig, 2019.

[4] Ani Nekova i Kathleen MvKeown. A survey of text summarization techniques. *Mining text data*, 43–76, Springer, Boston, MA, USA, 2012.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li i Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683*, octubre, 2019.

[6] Dragomir R. Radev, Eduard Hovy i Kathleen McKeown. Introduction to the special issue on summarization. *Computational linguistics*, 28.4:399–408, 2002.

[7] E. Cambria i B. White. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, 9:2:48–57, maig, 2014.

[8] Elizabeth D. Liddy *Natural Language Processing*. In Encyclopedia of Library and Information Science. Marcel Decker, Inc., NY, USA, segona edició, 2001.

[9] Frank Rosenblatt. The perceptron – A perceiving and recognizing automaton. *Report 85, Cornell Aeronautical Laboratory*, 460–461, 1957.

[10] Frederic Morin i Yoshua Bengio. Hierarchical probabilistic neural network language model. *In Proceedings of the international workshop on artificial intelligence and statistics*, 246–252, 2005.

[11] Ilya Sutskever, Oriol Vinyals i Quoc V. Le. Sequence to Sequence Learning with Neural Networks. *arXiv preprint arXiv:1409.3215*, setembre, 2014.

[12] Inderjeet Mani. Recent developments in text summarization. *The tenth international conference on Information and knowledge management (CIKM '01)*, 529–531, octubre, 2001.

[13] Informació del grup ELiRF. Consultat a https://www.dsic.upv.es/int/info.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee i Kristina Toutanova BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, octubre, 2018.

[15] Jeffrey L. Elman, Finding structure in time, *Cognitive Science*, 14:2: 179–211, 1990.

[16] Jingqing Zhang, Yao Zhao, Mohammad Saleh i Peter Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *Proceedings of the 37th International Conference on Machine Learning, PMLR*, 119:11328-11339, 2020.

[17] José Ángel González Barba. *Aprendizaje profundo para el procesamiento del lenguaje natural (tesis de master)*. Universitat politècnica de València, 2017

[18] José Ángel González, Encarna Segarra, Fernando García-Granada, Emilio Sanchís, Lluís Felip Hurtado. Extractive summarization using siamese hierarchical transformer encoders. *Journal of Intelligent & Fuzzy Systems Preprint*, 1–11, 2020.

[19] José Ángel González, Encarna Segarra, Fernando García-Granada, Emilio Sanchís, Lluís Felip Hurtado. Siamese hierarchical attention networks for extractive summarization. *Journal of Intelligent & Fuzzy Systems*, 36.5: 4599–4607, 2019.

[20] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman i Phil Blunsom Teaching machines to read and comprehend. *arXiv preprint arXiv:1506.03340*, juny, 2015.

[21] Keivan Kianmehr, et al. Text summarization techniques: SVM versus neural networks. *In: Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, 487–491, desembre, 2009.

[22] Liam ScanLon, et al. Evaluation of Cross Domain Text Summarization. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1853–1856, 2020.

[23] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang i Zhi Jin. How Transferable are Neural Networks in NLP Applications? *arXiv preprint arXiv:1603.06111*, març, 2016.

[24] Llistat d'aplicacions del PLN. Consultat a https://www.analyticsvidhya.com/blog/2020/07/top-10-applications-of-natural-language-processing-nlp/.

[25] Marc Moreno Lopez i Jugal Kalita. Deep Learning applied to NLP *arXiv preprint arXiv:1703.03091*, març, 2017.

[26] Mehdi Allahyari, et al. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*, juliol, 2017.

[27] Mike Lewis, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, octubre, 2019.

[28] Mitchell, R., J. Michalski i T. Carbonell. *Machine learning: an artificial intelligence approach*. Springer, Berlí, 2013.

[29] Naila Murray i Florent Perronnin. Generalized Max Pooling. *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2473–2480, 2014.

[30] Pedro Javier Ortíz Suárez, Benoît Sagot i Laurent Romany. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Leibniz-Institut für Deutsche Sprache, 2019.

[31] Piotr Bojanowski, Edouard Grave, Armand Joulin i Tomas Mikolov Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*, juliol, 2016.

[32] Prakash M Nadkarni, Lucila Ohno-Machado i Wendy W Chapman Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18:5:544–551, setembre, 2011.

[33] Rada Mihalcea i Paul Tarau. TextRank: Bringing Order into Text. *Proceedings of the 2004 conference on empirical methods in natural language processing*„ 404–411, 2004.

[34] Ritika Wason. Deep learning: Evolution and expansion. *Cognitive Systems Research*, 52:701–708, 2018.

[35] Saad Albawi, Tareq Abed Mohammed i Saad Al-Zawi. Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)*, 1–6, 2017.

[36] Stuart Russell i Peter Norving. *Artificial intelligence: a modern approach*. Pearson Education, Inc., Upper Saddle River, New Jersey 07458, USA, tercera edició, 2009.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado i Jeffrey Dean Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv:1310.4546*, octubre, 2013.

[38] Tomas Mikolov, Kai Chen, Greg Corrado i Jeffrey Dean Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*, gener, 2013.

[39] Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang i Min Sun. A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss. *arXiv preprint arXiv:1805.06266*, juliol, 2018.

[40] Yen-Chun Chen i Mohit Bansal. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. *arXiv preprint arXiv:1805.11080*, maig, 2018.

[41] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis i Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *arXiv preprint arXiv:2001.08210*, gener, 2020.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer i Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, juliol, 2019.

APÈNDIX A
Exemples de resum

A.1 Exemples de resums en català

Notícia

L ' entesa ha tardat a materialitzar-se , però la declaració de Mariano Rajoy en el cas Gürtel va propiciar ahir el primer acord de pes entre Pedro Sánchez i Pablo Iglesias des del retorn del secretari general socialista . Els dos líders van donar així el primer impuls a la taula de coordinació parlamentària creada fa menys de dues setmanes i que havia arrencat amb friccions . El PSOE i Podem van aliar-se ahir per convocar la reunió de la diputació permanent del Congrés amb un únic objectiu : la convocatòria d ' un ple extraordinari perquè el president espanyol " expliqui els motius pels quals es nega a assumir responsabilitats polítiques pels casos de corrupció que afecten el PP " i pels quals va haver de testificar a l ' Au- diència Nacional . Ara qui té l ' última paraula és la presidenta del Congrés , l ' exministra Ana Pastor (PP) , que decidirà quan es convoca la diputació permanent , la representació ponderada del ple en període de vacances . Podria deixar-ho fins a finals d ' agost , una forma de retardar la celebració d ' un ple extraordinari . Perquè Rajoy sigui cridat a donar explicacions , Podem i el PSOE necessiten sumar el suport d ' ERC i el PDECat -que es dona per segur- , així com el del PNB , que de moment ha dit que " no s ' hi oposarà " però no ha confirmat el sentit del seu vot . Ciutadans , en canvi , considera que el millor espai perquè Rajoy comparegui és a la comissió d ' investigació al Congrés sobre el finançament il·legal del PP . " És un lloc molt més concret i incòmode que un míting des de la tribuna del Congrés " , va assenyalar ahir Albert Rivera en declaracions des de la cambra baixa poc abans d ' anar a la Moncloa per assistir al primer acte públic del president espanyol després d ' haver testificat pel cas Gürtel , en què Rajoy va presumir de les bones dades econòmiques de l ' última EPA . Fins ahir , els socialistes opinaven el mateix que Ciutadans . De fet , Sánchez no va voler sumar-se dimecres a la petició de Podem de convocar un ple d ' urgència . " És una bona notícia que el PSOE s ' estigui acostant a nosaltres i que puguem col·laborar en la mateixa direcció " , va dir Iglesias en una entrevista a Efe . La portaveu dels socialistes al Congrés , Margarita Robles , va matisar la seva posició i va treure importància a la necessitat que el ple es faci com més aviat millor . Les posicions entre els dos partits continuen acostant-se per dalt , però no en els càrrecs intermedis .

Resum

Intenten forçar la compareixença urgent del líder popular al ple del Congrés.

Oracle de ROUGE

La portaveu dels socialistes al Congrés , Margarita Robles , va matisar la seva posició i va treure importància a la necessitat que el ple es faci com més aviat millor .

Oracle de BERTScore

Ara qui té l ' última paraula és la presidenta del Congrés , l ' exministra Ana Pastor (PP) , que decidirà quan es convoca la diputació permanent , la representació ponderada del ple en període de vacances .

Lead-2

L ' entesa ha tardat a materialitzar-se , però la declaració de Mariano Rajoy en el cas Gürtel va propiciar ahir el primer acord de pes entre Pedro Sánchez i Pablo Iglesias des del retorn del secretari general socialista . Els dos líders van donar així el primer impuls a la taula de coordinació parlamentària creada fa menys de dues setmanes i que havia arrencat amb friccions.

SHANN

Ara qui té l ' última paraula és la presidenta del Congrés , l ' exministra Ana Pastor (PP) , que decidirà quan es convoca la diputació permanent , la representació ponderada del ple en període de vacances . Ciutadans , en canvi , considera que el millor espai perquè Rajoy comparegui és a la comissió d ' investigació al Congrés sobre el finançament il·legal del PP.

mBART

El PSOE i Podem convoquen la reunió de la diputació permanent perquè el president espanyol “expliqui els motius pels quals es nega a assumir responsabilitats polítiques pels casos de corrupció”.

TextRank

['l ”', '”', 'ara qui té l ' última paraula és', 'congrés amb', 'un ple d ' urgència', 'faci com més aviat millor', 'com més aviat millor', 'expliqui els motius pels quals', 'els motius pels quals', 'ara qui té l']

A.1.1. Puntuacions

Sistema	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BERTScore
O. ROUGE	29,17	4,35	20,83	20,83	67,38
O. BERTScore	27,45	8,16	15,69	15,69	67,58
Lead-2	16,47	2,41	9,41	9,41	65,40
SHANN	22,78	5,19	15,19	17,72	68,73
mBART	8,51	0,00	8,51	8,51	67,21
TextRank	14,29	3,70	10,71	10,71	59,29

Taula A.1: Resultats dels diferents sistemes de generació de resums aplicats sobre la notícia d'exemple en català.

A.2 Exemples de resums en castellà

N̂otícia

El califato erigido por el grupo yihadista Daesh pierde terreno día a día en Siria e Irak , pero el grupo mantiene su capacidad de sembrar el terror como demostró en su primer gran ataque desde la pérdida a comienzos de julio de Mosul , la que fue su capital desde 2014 . Al menos 74 personas perdieron la vida y otras 93 resultaron heridas en un doble ataque perpetrado por los yihadistas cerca de Nasiriya , al sur de Irak . El objetivo de la operación , reivindicada por el grupo a través de la agencia Amaq , fue una estación de servicio en plena autopista y un puesto de control cercano . El primer ataque se produjo contra un restaurante frecuentado por peregrinos chiíes , que vienen de Irán para visitar las ciudades santas de Karbala y Nayaf , cuando un suicida se inmoló a la entrada y varios hombres abrieron fuego sobre los clientes . Poco después un kamikaze se inmoló en un puesto de control . Según los medios iraquíes , los autores iban vestidos con uniformes y distintivos de las Unidades de Movilización Popular , las milicias chiíes que combaten junto al Ejército y que han sido claves para la derrota de Daesh , y conducían también vehículos militares . Entre los fallecidos se confirmó la presencia de al menos diez ciudadanos de origen iraní , informaron fuentes del ministerio de Salud iraquí . Los expertos alertan de que la pérdida física del califato puede traducirse en un aumento de los atentados por parte de un grupo que en Irak solo conserva cuatro localidades .

Resum

Daesh ha reivindicado la autoría del ataque contra un restaurante y un puesto de control de la policía .

Oracle de ROUGE

Poco después un kamikaze se inmoló en un puesto de control .

Oracle de BERTScore

El objetivo de la operación , reivindicada por el grupo a través de la agencia Amaq , fue una estación de servicio en plena autopista y un puesto de control cercano .

Lead-2

El califato erigido por el grupo yihadista Daesh pierde terreno día a día en Siria e Irak , pero el grupo mantiene su capacidad de sembrar el terror como demostró en su primer gran ataque desde la pérdida a comienzos de julio de Mosul , la que fue su capital desde 2014 . Al menos 74 personas perdieron la vida y otras 93 resultaron heridas en un doble ataque perpetrado por los yihadistas cerca de Nasiriya , al sur de Irak .

SHANN

El califato erigido por el grupo yihadista Daesh pierde terreno día a día en Siria e Irak , pero el grupo mantiene su capacidad de sembrar el terror como demostró en su primer gran ataque desde la pérdida a comienzos de julio de Mosul , la que fue su capital desde 2014 . El primer ataque se produjo contra un restaurante frecuentado por peregrinos chiíes , que vienen de Irán para visitar las ciudades santas de Karbala y Nayaf , cuando un suicida se inmoló a la entrada y varios hombres abrieron fuego sobre los clientes .

mBART

Al menos 74 personas perdieron la vida en un doble ataque perpetrado por los yihadistas cerca de Nasiriya, al sur de Irak.

TextRank

['salud iraquí', 'agencia amaq', 'movilización popular', 'daesh', 'día', 'califato', 'irak', 'fuego', 'ministerio', 'su primer gran ataque']

A.2.1. Puntuacions

Sistema	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**ROUGE-Lsum**	**BERTScore**
O. ROUGE	31,25	20,00	31,25	31,25	68,27
O. BERTScore	34,62	20,00	26,92	26,92	73,33
Lead-2	20,00	0,00	14,00	14,00	63,74
SHANN	22,41	3,51	17,24	20,69	67,96
mBART	23,81	0,00	19,05	19,05	68.65
TextRank	15,79	0,00	15,79	15,79	61,22

Taula A.2: Resultats dels diferents sistemes de generació de resums aplicats sobre la notícia d'exemple en castellà.