UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dep. de Sistemes Informàtics i Computació

Resum abstractiu de notícies en català basat en models neuronals extractius i abstractius

Treball Fi de Màster

Màster Universitari en Intel·Ligència Artificial: Reconeixement de Formes i Imatge Digital

AUTOR/A: Marco García, Pere

Tutor/a: Segarra Soriano, Encarnación

Cotutor/a: Hurtado Oliver, Lluis Felip

CURS ACADÈMIC: 2021/2022

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, els sistemes de generació automàtica de resums suposen una gran ajuda per tal d'aconseguir resums adjunts i conéixer el contingut d'un document. En els sistemes basats en xarxes neuronals, els quals presenten els millors resultats en l'actualitat, existeix una limitació en la quantitat de text que poden rebre com entrada. Aquest fet pot generar perdudes d'informació en documents de gran extensió com pot ser el cas dels articles d'investigació o textos periodístics.

En aquest treball proposem utilitzar un model neuronal extractiu per tal de realitzar una selecció del contingut dels textos per a posteriorment aplicar un model de generació automàtica de resums abstractius. En aquest procés, s'utilitzaran sistemes basats en xarxes neuronals i s'aplicaran sobre un corpus de notícies en català on predominen els resums de referència de naturalesa abstractiva. Els resums generats pels models utilitzats s'avaluaran mitjançant mètriques de caràcter sintàctic i semàntic per poder comparar la qualitat dels resums.

Finalment, s'avaluaran els resultats obtinguts amb l'aplicació de la selecció de contingut i s'estudiarà la influència del preprocessat en els resums generats pels models neuronals abstractius.

Paraules clau: resum de textos periodístics, resum extractiu, resum abstractiu, transformers, català

Resumen

La cantidad de información de la que se dispone en la actualidad no ha hecho más que incrementar enormemente en las últimas décadas. En consecuencia, disponemos de un gran número de textos de los que tenemos que determinar su utilidad para cualquier actividad concreta. En este contexto, los sistemas de generación automática de resumen suponen una gran ayuda para conseguir resúmenes adjuntos y conocer el contenido de un documento. En los sistemas basados en redes neuronales, los cuales presentan los mejores resultados en la actualidad, existe una limitación en la cantidad de texto que pueden recibir en la entrada. Este hecho puede generar pérdidas de información en documentos de gran extensión, como puede ser el caso de artículos de investigación o textos periodísticos

En este trabajo proponemos utilizar un modelo neuronal extractivo para realizar una selección del contenido de los textos para posteriormente aplicar un modelo de generación automática de resúmenes abstractivos. En este proceso, se utilizarán sistemas basados en redes neuronales y se aplicarán sobre un corpus de noticias en catalán en el que predominan los resúmenes de referencia de naturaleza abstractiva. Los resúmenes generados por los modelos utilizados se evaluarán mediante métricas de carácter sintáctico y semántico para poder comparar la calidad de los resúmenes.

Por último, se evaluarán los resultados obtenidos con la aplicación de la selección de contenido y se estudiará la influencia del preprocesado en los resúmenes generados por los modelos neuronales abstractivos.

Palabras clave: resumen de textos periodísticos, resumen extractivo, resumen abstractivo, transformers, catalán

Abstract

The amount of information available today has only increased enormously in recent decades. Consequently, we have a large number of texts from which we have to determine their usefulness for any specific activity. In this context, the text summarization systems are a great help to get attached summaries and to know the content of a document. In systems based on neural networks, which present the best results at present, there is a limitation in the amount of text that they can receive in the input. This fact can generate loss of information in long documents, such as research articles or journalistic texts.

In this work we propose to use an extractive neural model to make a selection of the content of the texts to later apply an abstractive text summarization model. In this process, systems based on neural networks will be used and applied to a corpus of news in Catalan in which reference summaries of an abstract nature predominate. The summaries generated by the models used will be evaluated using syntactic and semantic metrics in order to compare the quality of the summaries.

Finally, the results obtained with the application of content selection will be evaluated and the influence of preprocessing on the summaries generated by abstractive neural models will be studied.

Key words: journalistic text summarization, extractive summarization, abstractive summarization, transformers, Catalan

Índex

Índex de figures

Índex de taules

CAPÍTOL 1
Introducció

Amb l'augment de la utilització d'internet i plataformes de serveis al núvol la quantitat de documents a la nostra disposició no fa més que augmentar d'una manera accelerada. Aqueste quantitats d'informació no només creixen en l'àmbit d'usuari per motius d'oci o passatemps, sinó que també involucren a quasi tots els àmbits professionals. Actualment és quasi obligatori dependre de la qualitat de recursos informàtics com són els buscadors per poder limitar el nombre de llocs en el que buscar informació. Tot i això en molts casos aquest filtratge no és suficient i la quantitat de documents a revisar fins a trobar la informació que necessitem pot fàcilment aplegar a ser inabastable en un temps raonable.

El resum automàtic de textos suposa una ajuda important en aquestes situacions quan a alleugerar càrrega de recerca es refereix. 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 [1]. Amb aquesta acció es pot col·laborar a agilitzar les tasques de tractament i aplicació d'informació en àmbits tan diversos com la bibliografia mèdica, els documents legals, els articles periodístics o les transcripcions de recursos audiovisuals. Hui en dia ha augmentat l'ús d'aquest tipus de sistemes, tendint a obtindre els millors resultats amb sistemes massius basats en xarxes neuronals i fent servir quantitats incommensurables de dades. Aquests sistemes són elaborats per grans corporacions i principalment dirigides al tractament d'arxius en anglés o multilingües. Amb aquesta tendència, institucions amb menor pressupost i llengües menys representades en documents informatitzats es troben a una situació de desavantatge. És per tot açò que l'ús de corpus i mètodes alternatius per a la millora de sistemes de generació automàtica de resums ha guanyat tanta rellevància recentment.

En el present treball ens centrarem en el preprocessament de textos per a la millora de prestacions a l'aplicar sistemes de generació de resums abstractius basats en xarxes neuronals com són NASCA, mBART i mT5. Per a aquest preprocessament utilitzarem sistemes extractius de generació automàtica de resums, com és el cas de SHANN. Per a la realització dels experiments farem servir un conjunt de notícies en català, que forma part del corpus bilingüe DACSA. Aquest corpus es compon per un elevat nombre de notícies que disposen 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. Tant DACSA com NASCA són recursos desenvolupats pel grup de recerca ELiRF, amb el qual s'hi ha col·laborat durant el transcurs d'una beca de l'institut VRAIN. Amb el corpus processat s'obtindrà un conjunt de frases de les notícies, de menor mida però amb l'objectiu d'obtindre la màxima informació rellevant possible. Finalment s'avaluarà mitjançant diferents mètriques les prestacions de diferents sistemes basats en xarxes neuronals enfront de l'entrada de textos compactats amb una estructura diferent.

1.1 Motivació

La intel·ligència artificial és un camp que busca comprendre i recrear el funcionament de la ment humana i com aquesta és capaç de percebre, interpretar i fins i tot predir els esdeveniments d'un món complex i aparentment indeterminista. Actualment segueix existint un debat amb la definició de la intel·ligència artificial. Les principals postures defenen que es tracta de l'emulació o recreació 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 la imitació de la capacitat creativa.

L'aprenentatge automàtic (*Machine Learning*) és una de les disciplines de la informàtica destinada al calc d'accions humanes, sobretot amb el propòsit d'obtindre els mateixos resultats. Les solucions desenvolupades amb l'aprenentatge automàtic es poden trobar en pràcticament tots els àmbits de la ciència i ha tingut impacte en la societat. Les tasques a les quals es pot aplicar varien enormement entre sectors de la societat incloent tasques de reconeixement d'imatges, emulació d'interaccions socials o anàlisi i extracció d'informació a partir de textos. [2].

Aquesta última disciplina forma part de l'àrea del Processament del Llenguatge Natural (PLN). El PLN consisteix en un ample ventall de tècniques computacionals dirigides a l'anàlisi i representació del llenguatge humà. Des del seu naixement a la dècada dels anys cinquanta, la recerca del PLN s'ha centrat en tasques com la classificació de textos, reconeixement de la parla i la cerca de respostes. Aquesta recerca al llarg dels anys s'ha centrat principalment en l'àmbit sintàctic, ja que era la forma més intuïtiva d'enfrontar-se amb els problemes a resoldre. Un enfocament semàntic en la resolució dels problemes del PLN sembla prometre millors resultats en la majoria de disciplines en les quals es treballa. Fins fa poc temps, les tècniques basades en l'anàlisi semàntic han quedat en segon pla a causa de la seua complexitat i la manca de grans volums d'informació etiquetada. [3].

Actualment, amb la popularitat de les xarxes neuronals i el desenvolupament de l'aprenentatge profund (*Deep Learning*), les aproximacions semàntiques d'aquesta àrea de recerca estan presentant millors resultats que les solucions basades en l'aproximació sintàctica i *Mavhine Learning* clàssic. En l'actualitat, amb representacions com els *transformers*, grans empreses i institucions tecnològiques han desenvolupat enormes xarxes neuronals especialitzades en el tractament i comprensió semàntica de textos per a la resolució de diferents problemes de PLN.

La creació, execució i anàlisi de ferramentes 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. Un exemple el trobem amb BART o GPT-3, sistemes d'utilitat general que són a més de gran importància en aquest camp i que ofereixen resultats prometedors per a la generació automàtica de textos. Així mateix, aquesta disciplina del PLN compta actualment amb un gran nombre de corpus i sistemes amb els quals poder treballar i dels quals podem obtindre resultats per poder realitzar una adequada recerca. Com que actualment els grans sistemes de generació automàtica de textos abasten tot el mercat, els grups de recerca amb recursos limitats tenim la necessitat i la responsabilitat d'abordar aquesta tasca de noves maneres que permeten obtindre resultats similars amb menor requisits computacionals.

També cal esmentar que els grans sistemes que es presenten en l'actualitat es desenvolupen utilitzant conjunts de dades en anglés i creant alternativament una versió multilingüe amb un nombre molt elevat de llengües. D'aquesta manera, les llengües minoritàries al món es troben novament en una situació de desavantatge al no veure explotat adequa-

dament els recursos que disposen per a la recerca. Habitualment les solucions en aquestes llengües presenten pitjors prestacions que els sistemes en anglés per no tindre models dedicats únicament a elles. La utilització de corpus de llengües minoritàries com el català ajuda, no només a la llengua a augmentar la seua representació i recursos, sino també als treballs de recerca per a explorar noves solucions i obtindre resultats competitius en àmbits en els quals falta molt per desenvolupar.

1.2 Objectius

En aquest treball tenim l'objectiu d'avaluar la influència de la selecció de contingut per a l'entrada d'un sistema de generació automàtica de resums. Amb aquest propòsit es pretén preprocessar un corpus de dades en català fent ús de sistemes extractius basats en xarxes neuronals i avaluar el seu funcionament a l'aplicar-se sobre sistemes de generació de resums abstractius. Amb aquesta finalitat podem distingir els següents treballs a realitzar:

1. **Anàlisi de les característiques del conjunt de dades.** S'estudiaran les qualitats del conjunt de notícies amb el qual s'ha decidit treballat amb la finalitat de conèixer el corpus i decidir el camí a seguir a la metodologia de l'experimentació.
2. **Processament de textos de notícies mitjançant sistemes de resum extractiu** S'aplicaran sistemes de selecció de frases basats en sistemes de resum extractiu amb la finalitat d'obtindre versions filtrades del conjunt de documents original.
3. **Utilització de sistemes basats en xarxes neuronals per a la generació de resums.** S'utilitzarà diferents models per a la generació automàtica de resums punters en aquesta tasca. Per a la seua utilització es consideraran diverses variants i parametritzacions.
4. **Avaluació dels 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,

1.3 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[1]. 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): Un dels principals objectius de la generació automàtica de resums és permetre el ràpid accés a grans quantitats d'informació extraient el contingut principal del text a resumir. D'aquesta manera amb la generació de resums, d'igual manera que ocorre amb la majoria d'àrees de recerca del PLN, ajudem a la qualitat de l'educació mitjançant facilitar l'accés a la informació, d'una manera o altra.

[1] Per a més informació es pot consultar el següent enllaç: https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/.

Figura 1.1: Aquests són els 17 objectius de desenvolupament sostenible establits per l'ONU a l'agenda per a 2030.

- **Energia assequible i no contaminant** (Objectiu 7): En aquest treball s'aborda la millora de l'eficiència energètica de dos punts de vista. Des del punt de vista la generació automàtica de resums aconseguim que no siga necessari duplicar tanta quantitat d'informació a enviar, emmagatzemar i tractar per saber la temàtica de les notícies. Des del punt de vista de l'aprenentatge automàtic, trobar maneres de millorar les prestacions dels models sense augmentar la mida de les xarxes neuronals pot tindre un gran impacte en el consum generat per la utilització d'aquest tipus de sistemes.

- **Reducció de les desigualtats** (Objectiu 10): En aquest treball fem ús tant del corpus DACSA com de NASca. DACSA, com explicarem més endavant, es tracta d'un dels corpus de text i de notícies més grans, tant en castellà com en català. NASca es tracta d'un sistema de generació automàtica de resums destinat a textos en català. Com és sabut, el català és una llengua que ha estat minoritzada durant molts anys, fins i tot a territoris on és la llengua majoritària. Amb la utilització d'aquests dos elements intentem fomentar la utilització d'aquesta llengua en els àmbits de recerca i propiciar el desenvolupament de sistemes destinats a treballar en aquesta llengua en l'àmbit del PLN.

1.4 Estructura de la memória

Per tal de donar al lector una visió global del treball, a continuació enumerarem les distintes parts d'aquesta memòria i descriurem breument els aspectes principals dels diferents apartats.

- **Capítol 1, Introducció.** En aquest primer capítol s'explica quina ha sigut la motivació del projecte, l'àmbit en el qual s'ha desenvolupat, els principals objectius que es persegueixen i la seua finalitat del treball.

- **Capítol 2, Estat de la qüestió.** En el segon capítol es pretén contextualitzar el treball per donar una visió més concreta d'aquesta branca de l'aprenentatge automàtic. Amb aquesta finalitat es descriu la situació actual de la generació automàtica de resums i es presente els principals sistemes que s'utilitzen hui en dia per a aquesta tasca.

- **Capítol 3, Marc teòric.** En aquest capítol es descriuen els sistemes de representació de textos, les mètriques d'avaluació i els mètodes de preprocessat de textos utilitzats en el treball. A més, s'exposaran els diferents sistemes de generació de resums basats en xarxes neuronals utilitzats al llarg del projecte i les modificacions realitzades sobre els sistemes i el corpus.

- **Capítol 4, Ferramentes utilitzades.** En el capítol quart s'explica les diferents plataformes, hardware i ferramentes que han sigut necessàries per a la utilització i modificació dels diferents elements utilitzats en el treball.

- **Capítol 5, Experimentació i resultats.** En el cinqué capítol es mostra els experiments portats a terme amb els diferents models descrits anteriorment utilitzant diverses versions del corpus i es presenten els valors obtinguts per les mètriques aplicades sobre els resums obtinguts.

- **Capítol 6, Conclusions.** A l'últim capítol exposem quines han sigut les conclusions que s'ha pogut extraure dels resultats obtinguts i del treball desenvolupat. També es presentaran les possibles vies de treball amb les quals continuar a partir dels resultats del present treball.

1.5 Context i col·laboracions

La realització d'aquest projecte s'ha realitzat durant el gaudiment d'una beca formativa de Col·laboració en l'Institut Valencià d'Intel·ligència Artificial (VRAIN[2]) a la Universitat Politècnica de València (UPV).

Aquestes beques tenen la finalitat de formar graduats universitaris en el món de la recerca en l'àmbit informàtic treballant conjuntament amb grups de recerca de l'institut en casos d'interés reals i relacionats amb la intel·ligència artificial. En el nostre cas concret s'ha realitzat la Col·laboració amb el grup ELiRF[3] (Enginyeria del Llenguatge i Reconeixement de Formes), grup de recerca especialitzat amb l'àrea del PNL. El projecte concret de col·laboració ha sigut el resum abstractiu de textos basats en xarxes neuronals on es pretén utilitzar els models de l'estat de l'art sobre el corpus DACSA i l'aplicació de mètriques usuals per a l'avaluació dels resultats.

Durant la realització de la beca s'ha comptat amb l'ajuda i consell d'Encarna Segarra Soriano, Lluís Felip Hurtado Oliver, José Ángel González Barba i Vicent Ahuir Esteve. També s'ha utilitzat materials propis del grup com sistemes de generació automàtica de resums basats en xarxes neuronals com són SHA-NN i NASca el corpus DACSA i models dels sistemes que componen l'estat de l'art ajustats per a la tasca de resums sobre aquest corpus concret. Per últim, aquesta col·laboració també ha facilitat disposar de recursos específics essencials en determinades labors, necessàries per a l'adequada realització del projecte.

1.6 Relació amb els estudis cursats

El desenvolupament del present treball s'ha vist influenciat per diverses de les assignatures cursades al Màster Universitari en Intel·ligència Artificial, Reconeixement de Formes i Imatge Digital. Les principals assignatures que han aportat coneixements a aquest treball

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

[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.

han sigut les relacionades amb la branca de Tecnologies del Llenguatge com són Lingüística Computacional (LC), Reconeixement Automàtic de la Parla (RAH) i Aplicacions de la Lingüística Computacional (ALC).

Aquestes assignatures han contribuït amb informació necessària per a conéixer l'actualitat del processament del llenguatge natural, els sistemes de representació adequats i les diferents tècniques i sistemes que formen part de l'estat de l'art d'aquesta branca de recerca.

Altres assignatures del màster cursat han contribuït en diferents àrees aportant coneixements sobre la utilització de models basats en xarxes neuronals, millorant les possibilitats i característiques del llenguatge de programació Python i proporcionant experiència i fluïdesa a l'hora de poder interpretar els resultats de l'experimentació a través del conjunt de treballs realitzats a les diferents branques dels estudis realitzats.

CAPÍTOL 2
Estat de la qüestió

En el present capítol presentarem la importància del Processament del Llenguatge Natural i aprofundirem en la generació automàtica de resums i el seu estat actual. També presentarem la situació dels corpus en el present i exposarem la tendència actual a la intel·ligència artificial a presentar models cada vegada més pesats amb creixents requeriments a l'entrenament.

2.1 Processament del llenguatge Natural

Des de la seua aparició com àrea de la informàtica, el processament del Llenguatge Natural ha sigut una àrea molt important i activa de la intel·ligència artificial. La seua versatilitat per a col·laborar en els àmbits de la societat li ha proporcionat un gran interés per part del món de la recerca. Tot i això, aquesta disciplina mai ha deixat d'evolucionar i no existeix una definició estàndard d'allò que abasta. Elizabeth D. Liddy en [4] ens presenta 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.*

En aquesta definició observem clarament perquè aquesta disciplina pertany a la Intel·ligència Artificial. Com sabem existeixen dues maneres d'obtindre resultats similars als generats per un humà, replicant el procés que realitza el cos humà per a realitzar una tasca o produint, mitjançant vies alternatives, un sistema que obtinga els mateixos resultats, sense assegurar que l'enteniment del procés siga igual que el d'una persona. A causa de la complexitat de les tasques relacionades amb el tractament del llenguatge, en l'actualitat no s'ha aplegat a fer que les màquines aconseguisquen el mateix enteniment del llenguatge que l'esser humà. Aquesta branca de recerca es coneix com a NLU (*Natural Language Understanding*).

En l'actualitat, el processament del llenguatge natural el podem trobar en nombroses àrees de la societat. El podem trobar clarament en àmbits quotidians en l'ús de motors de cerca, correctors en editors de text o els predictors en xats. També el trobem oralment en assistents de veu, que clarament estan guanyant gran importància en els darrers anys i permeten realitzar cada vegada una major quantitat d'accions i més variades. En altres àmbits més delicats també presenta funcions rellevants, com en el resum de textos legals o en la detecció de paraules clau en informes mèdics per a l'assistència en el diagnòstic.

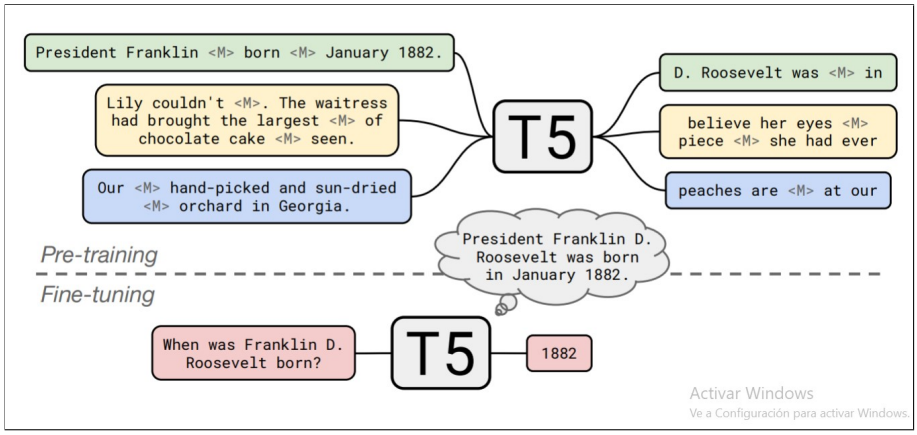

Figura 2.1: Representació visual del sistema T5 sent utilitzat per a resoldre diferents tasques del PLN [5].

A banda de sistemes aïllats que resolen de manera separada totes aquestes tasques, amb la utilització de xarxes neuronals ha augmentat la tendència de sistemes capaços de resoldre de manera acceptable diverses tasques del PLN alhora. Un exemple d'aquest fenomen el trobem amb GLUE [6], SuperGLUE [7] i T5 (*Text-to-Text Transfer Transformer*) [8]. 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 [9].

2.2 Generació automàtica de resums

Hui dia, la major disponibilitat de textos que permeten un millor entrenament dels models també implica una major necessitat per a trobar informació concreta. Tant en textos periodístics com en àmbits legals o de recerca, agilitzar el temps de lectura o conéixer breument el contingut d'un document té una gran importància. És per aquest motiu que en l'actualitat la generació automàtica de resums no para de guanyar importància. Aquesta tasca del PLN queda ben representada per la definició que ofereix Radev:

> *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 [10].*

Des que som capaços de fer ús del llenguatge, els humans som capaços de descriure els esdeveniments que ens envolten o que sentim per mitjà de gestos o el llenguatge oral. Aquestes descripcions solen ser incompletes, obviant informació i eliminat aquella que no n'és rellevant. D'igual manera, en l'escriptura ens resulta natural abreujar un document mantenint la informació mínima i necessària. Aquesta habilitat de les persones per a fer resums és el que en la generació automàtica de resums es pretén transmetre a les màquines per que siguen capaces d'entendre el contingut d'un document de text i extraure de manera coherent les idees principals. Com no coneixem tampoc quina és l'activitat que realitza el nostre cervell per a l'elaboració d'aquests resums els models que es desenvolupen intenten imitar el resultat mitjançant altres procediments que busquen aproximar el resultat.

Per poder saber que és el que s'espera dels sistemes desenvolupats en la generació automàtica de resums primer hem de diferenciar els diferents tipus de resum que podem obtindre. En primer lloc cal distingir entre resums extractius i resum abstractius. Els resums extractius són aquells que estan formats per fragments que es presenten en el mateix text original mentre que els resums abstractius són aquells que mantenen les idees del document sense necessitat de copiar les estructures literalment de la font.

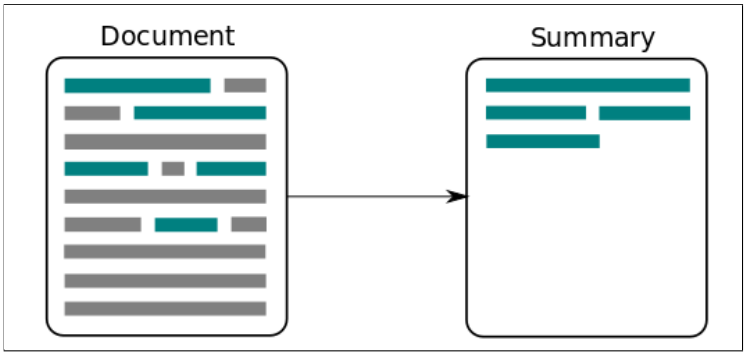

Figura 2.2: Exemple gràfic de la definició de resum. Interpretant que els fragments remarcats són els mateixos que es reordenen per a ser el resum, es tracta d'un exemple de resum extractiu

Una altra manera de catalogar els resums és pel nombre de documents a partir dels quals es genera. En la majoria de treballs, d'igual manera que ocorre en aquest projecte, els models es centren en la generació de resums a partir d'un únic document. No obstant, també existeixen treballs dirigits a resumir un conjunt de documents per tal de generar un únic resum conjunt. Aquest tipus de treball pot ser de gran utilitat per tal de classificar textos o alleugerar encara més la càrrega de treball necessària per a trobar informació concreta en la gran quantitat d'informació de la qual disposem hui en dia, però al mateix temps es veu augmentada la dificultat i el nombre de dades amb què treballar és significativament menor. Per últim, 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 [1].

Per a la resolució d'aquesta tasca, la majoria de solucions que es desenvolupen són de caràcter extractiu. Les primeres aproximacions buscaven paraules clau en el text que indicaren importància per al contingut del mateix d'acord amb el títol, altres es basen en generacions de grafs per a saber quins són els fragments als quals es fa més referència dins d'un mateix text, i altres, com Lead-K simplement seleccionen les K primeres línies del document per a generar el resum d'aquest. Per molt simple que parega aquest últim presenta una gran importància als noticiaris, per aquest motiu l'hem fet servir en el present treball i l'explicarem més endavant en la memòria. Hui en dia també es generen models extractius per a la generació automàtica de textos, com és el cas de l'arquitectura SHANN.

Com que el resum extractiu és més directe i més explotat hi ha un gran nombre de treballs que mesclen la generació extractiva amb l'abstractiva. Principalment podem distingir entre dos corrents principals, el *Reinforcement Learning model* (RL) [11] i el *Inconsistency Loss model* (IL) [12]. EL model RL es basa en l'entrenament per reforç de dos mòduls separats. En primer lloc s'entrena el model perquè seleccione mitjançant extracció les millors frases del document. Posteriorment la part abstractiva s'encarrega de comprimir el fragment per tal d'aconseguir major semblança amb el resum de referència. En el cas del IL l'estratègia consisteix a realitzar l'entrenament de les dues parts de manera conjunta. Tot i que l'estructura de l'entrenament és similar a RL les parts es centren en aspectes diferents. Mentre que la part extractiva es centra en l'àmbit de frase, la part abstractiva

busca entrenar a escala de paraula. 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 [13].

Actualment, el focus de la recerca també s'està dirigint als resums abstractius, ja que els seus resultats resulten més interessants i s'aproximen més al comportament humà en aquesta tasca. Actualment trobem diversos sistemes que, tot i que conserven una part extractiva, generen el resum de manera abstractiva. Alguns d'aquests exemples són l'arquitectura PEGASUS [14], BERT [15], i NASCA [16], el qual fem servir en el present treball i explicarem en la metodologia.

Amb l'evolució de les xarxes neuronals, d'igual manera que passa en altres disciplines de la IA, molts dels models que es publiquen recentment tenen la seua base en les xarxes neuronals. L'aprenentatge profund ha demostrat que té la capacitat de millorar les prestacions dels sistemes, no necessàriament considerant temps i memòria necessària, en moltes àrees de recerca. En PLN no trobem una excepció. Amb els transformers i les noves tecnologies de transfer learning les xarxes neuronals presenten resultats molt prometedors i que sembla que cal explorar més profundament per a extraure tot el seu potencial.

Tot i això, cal no oblidar la importància i la rellevància d'altres tecnologies tradicionals. En l'actualitat també trobem treballs que aborden la tasca de la generació automàtica de resums amb tecnologies conegudes com són Clustering, basats en grafs, SVM i Fuzzy Logic. Aquests treballs són importants, no només per les seues aportacions, sinó també per evitar la monotonia de les publicacions i explorar opcions diferents de la tendència general, cosa que en el món de la IA és sabut que pot arribar a aportar enormes beneficis.

2.3 Corpus

En l'actualitat, la disponibilitat de grans quantitats de dades no és condició suficient per a obtindre bons resultats en sistemes d'intel·ligència artificial. En la majoria de casos és determinant disposar d'un conjunt de dades de qualitat. Desafortunadament els corpus són difícils i cars d'aconseguir, sobretot en l'àrea del PLN. Per obtindre dades de gran qualitat i de gran mida és necessària la intervenció de moltes persones que etiqueten les dades i comproven la certesa d'aquestes. Aquesta feina costa molt temps i diners i sol estar portada a terme per grans empreses o institucions governamentals. L'alternativa a aquesta inversió és desenvolupar noves formes d'adquisició de dades, per exemple a través de la web, i moltes vegades s'ha de confiar en el correcte funcionament i filtratge de les dades.

És per això, entre altres motius, que la principal tendència en l'actualitat és la utilització de corpus en anglés generats per altres. A més, aquest tipus de dades presenten l'avantatge que, al popularitzar-se, sabem que han sigut revisats i utilitzats sobre altres sistemes, fent que els nous models siguen comparables amb altres treballs. Un exemple el trobem amb CNN/Daily Mail [17] i NewsRoom [18], *datasets* amb els quals es treballa per a crear la major part de les ferramentes de generació automàtica de resums. El corpus CNN/Daily Mail és un corpus d'articles de notícies format per prop de 300 000 parells únics de notícies i resums procedents dels diaris CNN i *Daily Mail*. NewsRoom, per la seua banda, està compost per 1,3 milions d'articles i notícies escrites pels autors i editors de prop de 40 mitjans de comunicació extrets dels seus llocs web. Amb aquesta mateixa metodologia es va crear el corpus DACSA, un corpus de notícies en català i en castellà, que hem fet servir en aquest projecte i que explicarem posteriorment.

2.4 Grans models

Amb l'èxit de les xarxes neuronals, s'ha vist en diverses ocasions la tendència a afegir més i més capes per tal d'aconseguir millors resultats. Tot i que s'ha desenvolupat tecnologies i alternatives perquè aquest creixement siga més eficient en l'actualitat aquesta tendència no s'ha aturat. Tot i que en diversos camps de la IA es presenta el mateix fenomen, en el PLN tenim un dels exemples més recents i més representatius. Quan en 2020 Open-AI va publicar GPT-3 molta gent es va sorprendre dels bons resultats presentats i l'aparent enteniment del llenguatge humà que suggerixen les respostes generades.

A partir d'aquest punt, els competidors de l'empresa han decidit optar per models de llenguatge massius arreu del món com són PanGu de Huawei o HyperCLOVA de Naver. Aquesta tendència, encara que presenta molt bons resultats, deixa de ser accessible no només per al seu entrenament, sinó per a la seua utilització per a la gran majoria de particulars i equips de recerca del món. Per aquest motiu, entre altres, és important que els equips de recerca siguen conscients d'aquest fenomen i que busquen o promoguen alternatives assequibles que aconseguisquen millors resultats amb noves tècniques i no únicament per acumulació de milions de paràmetres.

CAPÍTOL 3
Metodologia

3.1 Representació de documents

En el PLN, la representació dels textos és fonamental per al disseny del sistema i és determinant per a l'aprenentatge i prestacions dels models, independentment de la tasca a tractar. Tradicionalment s'ha fet servir les representacions de *One-Hot* i *Bag of Words* i les seues variants per al tractament de textos. Hui en dia en el PLN, d'igual manera que ha començat a passar en altres branques de la Intel·ligència Artificial, altra representació és la que predomina en els sistemes que presenten els millors resultats, els *embeddings*. A continuació explicarem les diferents formes de representació esmentades, els principals aspectes de cada una i de quina manera s'han fet servir els *embeddings* en el present treball.

3.1.1. One-Hot

El sistema de representació més directe per a passar un text a forma vectorial que es sol utilitzar en el PLN és el One-Hot, en aquesta representació cada paraula d'una frase es transforma a un vector de talla del vocabulari, n, tal que $x = \{x_1, x_2, ..., x_n\}$, on el vector pren el valor 1 en la posició que ocupa aquesta paraula al vocabulari i un 0 a la resta. D'aquesta manera, 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. La descripció formal de la creació d'aquest vector és la següent [19]:

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

	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.

Tot i que es tracta d'un sistema simple, immediat i fàcil d'implementar presenta una sèrie de problemes de notòria importància. En primer lloc, com hem comentat abans, la representació d'un text depén tant de la mida del text com de la grandària del vocabulari, obtenint en casos reals una matriu dispersa d'una dimensió molt ineficient. A més a més, obliga a mantindre una dimensionalitat fixa i no aporta informació sobre la semàntica del document.

3.1.2. Bossa de paraules

Una representació més avançada que la que acabem de descriure és Bossa de paraules o *Bag of words*. En aquest mètode, es crea un vector del mida del vocabulari per cada document. En la idea original de la Bossa de paraules aquest vector conté el número de vegades que cada paraula ha aparegut al document. Aquest vector es pot generar a partir de la representació de One-Hot sumant les files de la matriu per a cada document.

	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.

Una versió més complexa i utilitzada és la representació tf*idf (*Term Frequency*Inverse Document Frequency*) Sobre un conjunt de documents, 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. La descripció de les fórmules per al càlcul d'aquests valors són les següents:

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

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

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

En aquesta representació se li dona una major relevància a la freqüència relativa de les paraules tot i que es perd l'ordre d'aparició de cada token. També la representació és més eficient espacialment que en One-hot. Per solucionar el problema de l'ordenació existeix l'opció de realitzar el càlcul amb n-grames. També existeixen altres representacions que utilitzen altres mètodes de comptabilització i suavitzat dels resultats.

3.1.3. 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.

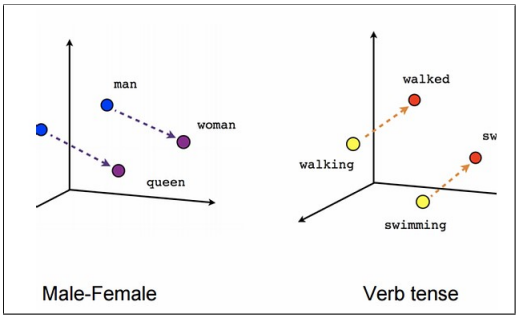

Figura 3.3: S'observa com parelles de paraules relacionades semànticament de la mateixa manera presenten una disposició similar.

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 [20].

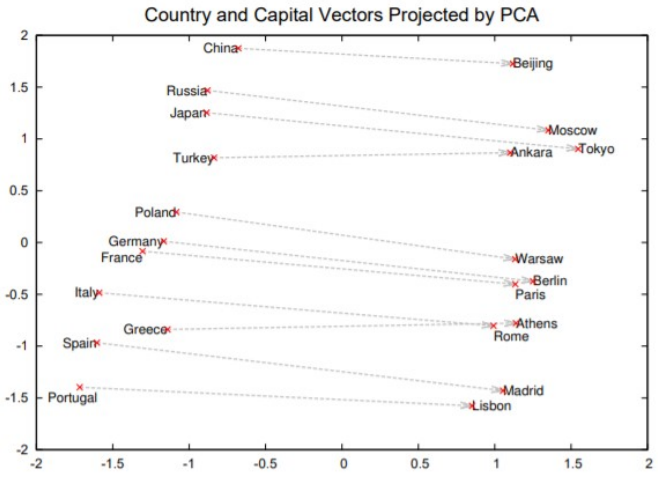

Figura 3.4: Representació mitjançant PCA de la distribució en *embeddings* dels noms de països i les seues capitals [20].

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.[1]

Per als models Word2Vec, per als quals cada paraula té assignat un vector, s'usa el model d'entrenament proposat per Mikolov, et al [22]. La finalitat de l'arquitectura *Skip-*

[1]Per trobar més informació sobre aquests models es recomana la referència [23].

gram és 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:

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

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*oftmax* 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 [21].

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 [15]. 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.

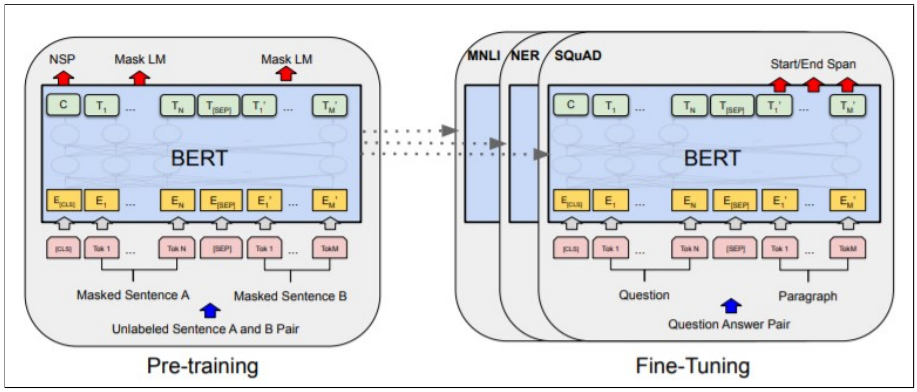

Figura 3.5: Representació de les etapes de *pre-training* i *fine-tuning* que realitza BERT per a l'aprenentatge dels seus paràmetres [15].

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.

3.2 Corpus utilitzats

3.2.1. Corpus DACSA

Com s'ha comentat amb anterioritat, el conjunt de dades que s'ha utilitzat per a l'obtenció dels resultats prové del corpus DACSA [24]. Aquest corpus consisteix, tant en català com en castellà, en parells de textos del tipus notícia-resum, on el resums han sigut escrits pels propis escriptors o editors dels mitjans de comunicació. Les notícies de DACSA s'ha extret seguint la metodologia del corpus de NewsRoom, és a dir, les notícies han sigut extretes de les pròpies webs periodístiques. La distribució de les notícies és la que podem veure a continuació, amb un total de 1 000 000 de notícies per a català i prop de 5 500 000 documents en castellà.

El conjunt de dades obtingut de les pròpies webs ha sigut depurat a través d'una sèrie de filtres per tal de millorar la qualitat i les característiques del corpus. Entre aquests filtres trobem l'establiment d'una llargària mínima per a les notícies i la comprovació de la similitud dels resums amb els inicis de les notícies. Les notícies que no han passat els diferents filtres entren en la categoria d'excloses. Després del filtratge les notícies incloses finalment superen les 725 000 per al conjunt en català i les 2 100 000 per al conjunt en castellà. Com que en aquest treball hem realitzat un estudi específic per al corpus en català, les dades i avaluacions que presentarem seran específiques per a aquesta llengua. A la taula 3.1 podem veure el desglossament per font del nombre d'articles que s'ha considerat per a la creació del corpus.

Font Periodística	**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.

Les notícies que han quedat incloses han sigut separades en 4 particions: **training**; 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); $\textbf{test}_V$, conjunt de notícies per a provar el funcionament de sistemes de tractament de textos; i $\textbf{test}_{NV}$, Aquesta partició és un conjunt distint de fonts que no vistes en cap etapa de l'entrenament ni el testege. Les fonts que s'han separat ha sigut amb motiu del baix nombre de notícies que presentaven. 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 testege cada una. La distribució per al català d'aquestes particions queda com podem observar a la taula 3.2.

Idioma	Training	Validation	Test_V	Test_{NV}
Català	636 596	35 376	35 376	17 836

Taula 3.2: Distribució del conjunt de notícies en les seues 4 particions per al català.

Per últim assenyalar la gran majoria de corpus que es disposa en l'actualitat, ja siga per a la generació automàtica de resums o la resta de tasques del PLN, es troben en anglés o l'inclouen principalment en tasques que requereixen més d'una llengua. Per als casos del castellà i el català es sol disposar d'un nombre molt reduït de mostres per a l'entrenament de models, o les seues publicacions tenen un menor impacte. En el cas del català concretament els principals corpus provenen d'articles de la Viquipèdia o el Corpus Oscar [25], que és el que Oscar, com és el cas del sistema mBART. Tenint açò en compte destaquem la importància de fer ús d'aquest corpus en aquest projecte i promoure el seu ús per a la realització de tasques de PLN en els àmbits que siga possible.

3.2.2. Corpus per al pre-entrenament

Donat que no és una bona pràctica realitzar el preentrenament de sistemes amb el mateix corpus que amb el que es va a realitzar l'experimentació, s'ha decidit seleccionar altres fonts, tant per a l'entrenament dels embedding com per al preentrenament dels models de NASca.

- **Oscar**: Com expliquen els seus autors, Oscar (*Open Super-large Crawled Aggregated coRpus*) es tracta d'un gran corpus multilingüe a partir de la classificació i filtratge del corpus *Common Crawl*. En el nostre cas farem servir els documents en català, en la seua versió sense mesclar i sense duplicats.
- **Viquipèdia** S'utilitzarà un bolcat de les entrades en català de Viquipèdia.
- **DACSA** Tant les notícies excloses per no passar els corpus, com el conjunt de notícies destinades a l'entrenament es faran servir per a l'entrenament dels models d'embeddings i el preentrenament de NASca.

3.3 Mètriques d'avaluació

Per poder avaluar els resums generats pels diferents models utilitzats a l'experimentació serà necessari seleccionar mètriques que permeten comparar diferents característiques dels resultats. EN el nostre cas hem seleccionat dues mètriques altament conegudes i utilitzades, les quals són Rouge i BertScore. La mesura de Rouge permet avaluar diferents aspectes relacionats amb l'estructura sintàctica dels textos, mentre que la mesura de BertScore guarda relació amb l'àmbit semàntic.

Per a la utilització d'aquestes mètriques, com s'explica en el capítol 4 d'aquesta memòria, s'ha fet servir la llibreria *Datasets* de *Hugging Face*. L'ús d'aquesta llibreria permet saber que la implementació és comuna a la utilitzada en altres treballs relacionats amb la mateixa tasca, i d'aquesta manera saber que els resultats obtinguts són fiables de comparar.

3.3.1. ROUGE

Donades dues cadenes de text, la mesura de ROUGE retorna un valor entre 0 i 1 segons el solapament que existeix entre les dues entrades. Generalment el valor de la cadena

només serà 1 si les dues cadenes són idèntiques i 0 si no tenen cap token en comú. Segons la variant de ROUGE que es faça servir les comparacions del solapament es realitza amb algunes diferències amb dues corrents principals, ROUGE-n i ROUGE-L. En les mesures del tipus ROUGE-N es té en compte el solapament de N-grames, mentre que amb ROUGE-L es consideren tots els elements dels textos. A continuació explicarem com es calculen els casos particulars d'estes dues aproximacions que s'han fet servir en el present treball.

ROUGE-1

Aquesta és la principal versió de ROUGE-N, la qual consisteix en la comparació de totes les paraules d'ambdues entrades. El seu càlcul consisteix a obtindre la superposició entre el resum generat i el resum de referència, típicament obtenint tres valors, *recall*, *precisió* i *f-score*. A continuació explicarem en detall el càlcul de cada un d'aquests valors, el qual es pot generalitzar per a les altres variants de ROUGE, ja que es tracta del cas més senzill.

En el cas de ROUGE-1, el *recall* es defineix com el nombre de paraules del resum de referència que es troben en el resum generat. La definició formal d'aquest enunciat s'expressa de la següent manera:

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

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.

Un exemple pràctic d'aquesta fórmula seria el que segueix. Considerant com a text generat S_g = "*El gat negre està baix del llit*ï S_r = "*El gat està baix del llit*çom a text de referència. Per a aquest cas observem que totes les paraules del resum de referència es troben al resum generat. Per tant el resultat queda com $Recall = \frac{6}{6} = 1,0$.

Altrament, la precisió es centra en el percentatge de paraules del text generat que són rellevants i ofereix una idea del fragment de text generat que no ofereix informació (segons el contingut literal del text de referència). Aquest càlcul es realitza comptant el nombre de paraules del resum generat que apareixen en el resum de referència, que seguint la nomenclatura de la fòrmula anterior queda definit de la forma següent:

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

Tenint en compte les mateixes frases presentades a l'anterior exemple, observem que la mida del text generat és de 7 paraules, mentre que només 6 d'aquestes es troben en el resum de referència. Com a resultat obtenim que el valor de la *precisió* d'aquest exemple és de $\frac{6}{7} = 0,86$.

Per últim, el valor *f-score* es tracta d'una relació entre els dos valors anteriorment descrits. Aplicant un escalar es combina la v i el *recall* on se li pot atorgar major rellevància a qualsevol dels dos, segons siguen els requeriments del problema o l'experiment. La fórmula per al càlcul de *f-score* s'expressa 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 $R(S_g, S_r)$ retorna el valor de *recall* entre el text generat i el de referència, la funció $P(S_g, S_r)$ obté el valor precisió entre el resum de referència i el resum generat, i β controla la importància relativa de les dues mesures.

En aquest treball l'aparició de totes les paraules del resum de referència al resum generat no s'ha considerat prioritari. D'igual manera, tampoc s'ha centrat l'atenció en els resums que només presenten paraules que sí que estan als resums aportats pel corpus. Per tot açò ens hem centrat en estudiar els resultats del valor *f-score* amb 1 com a valor per a la β, que ens ofereix una visió més general.

ROUGE-2

Com hem avançat, els valors que s'obtenen amb ROUGE-2 són els mateixos que amb ROUGE-1, amb l'excepció que el compteig es realitza amb parells de paraules en lloc d'amb paraules individuals. Per obtindre aquests valors es crea un vocabulari de bigrames de les frases que es coparen i es realitzen els sumatoris per a cada un d'aquests parells de paraules. Aquesta mesura ha mostrat estar relacionada amb la llegibilitat de les frases, per tant és important per saber que els resums generats tinguen certa proximitat a les frases creades per persones reals.

ROUGE-L

En el cas de ROUGE-L, a diferència dels casos de ROUGE-N, no es troba limitat a una mida determinat de n-grames, en el seu lloc realitza el càlcul basant-se en la cadena coincident de major longitud. En molts casos aquesta mesura es sol considerar la que millor determina la semblança entre els dos textos que es vol comparar.

3.3.2. BERTScore

BERTScore és un mètode automàtic d'avaluació per a tasques de generació de text. El seu funcionament es basa a obtindre 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.

Tot i que els valors teòricament sí que poden anar fins de 0 a 1, a diferència de ROUGE els valors obtinguts no calculen el percentatge de similitud entre dos textos. En aquest cas textos totalment diferents poden obtindre puntuacions superiors a $0,5$ mentre que textos pràcticament iguals poden no aplegar a $0,9$. Habitualment en català es solen obtindre valors propers a $0,65$ en els pitjors casos i de l'ordre de $0,85$ per a textos molt similars. A causa d'aquest fet és difícil saber quina és realment la diferència de similitud semàntica entre dos textos basant-nos amb el valor de BERTScore. Tot i això, sí que ens permet extraure que un text és més similar que un altre mitjançant comparar les puntuacions obtingudes per BERTScore.

A continuació presentem uns exemples de casos extrems on es pot comprovar el que s'ha explicat al paràgraf anterior. En el primer d'ells executem la instrucció:

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

Aquestes frases no tenen cap relació aparentment i presenten una estructura sintàctica i semàntica improbable. No obstat això, BERTScore ens dona una puntuació de $0,655962$.

En el següent cas presentem el cas contrari, on es comparen dues frases molt similars amb un significat molt proper:

```
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 ser pràcticament sinònimes, la seua puntuació és de $0,841035$.

3.4 Sistemes d'avaluació tradicionals

Tot i que aquest treball té l'objectiu de comparar els resultats de la generació automàtica de resums a partir de models basats en xarxes neuronals també s'ha fet ús de sistemes tradicionals. Les tècniques tradicionals són fonamentals per a algunes labors i són capaços d'obtindre bons resultats en alguns casos com desenvoluparem a continuació vorem a continuació. També permeten analitzar el comportament d'altres sistemes i obtindre unes bases amb les quals comparar resultats.

3.4.1. Lead-K

Els resums obtinguts mitjançant Lead-K s'obtenen extraient literalment les K primeres frases del text original. Aquest mètode presenta un gran avantatge respecte a altres mètodes tant en simplicitat d'implementació com en temps d'obtenció. L'aproximació més popular d'aquest mètode és Lead-3. Tot i això, les aproximacions Lead-1 i Lead-2 tenen gran rellevància en el context d'aquest treball per diferents motius. En primer lloc, en el corpus en què treballem els resums de les notícies presenten una longitud mitjana d'entre 1 i 2 (segon la font i la partició. A més, a causa de l'estructura intrínseca dels articles periodístics, la principal informació de la notícia apareix a les primeres frases del text, d'aquesta manera, en molts casos el resum de referència i el generat mitjançant Lead-K presente molts aspectes en comú.

3.4.2. Oracle

Quan parlem d'un oracle ens referim a un sistema que obtén el millor resum extractiu possible. El seu funcionament es basa a obtindre 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.

En aquest treball hem utilitzat un oracle que 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:

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

En el present treball hem utilitzat aquest oracle per poder realitzar una anàlisi més exhaustiu dels resums obtinguts.

3.5 Sistemes de resum extractiu basats en xarxes neuronals

Com hem explicat, els sistemes extractius són els que generen el resum a partir de fragments explícits del document original. En aquest treball, hem fet ús del sistema extractiu SHANN per al preprocessament de les notícies per a utilitzar posteriorment models abstractius.

3.5.1. SHANN

SAHNN [26], o *Siamese Hierarchical Neural Networks* és un sistema per a la generació automàtica de resums extractius 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.

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.

A l'hora de realitzar els resums d'un text, SHANN atorga una puntuació a cada una de les frases a partir de la qual genera el resum. En el nostre treball hem modificat el comportament de SHANN per tal que torne en lloc del resum, les puntuacions ordenades per ordre d'aparició en el text original (fins a un límit de 200 frases. Amb aquestes puntuacions hem obtingut les millors frases, mantenint l'ordre original, fins a completar un nombre determinat de tokens.

3.6 Sistemes de resum abstractiu basats en xarxes neuronals

Per a la generació automàtica de resums del corpus DACSA s'ha decidit fer ús dels models abstractius mBART, mT5 i NASca, els quals explicarem a continuació:

3.6.1. mBART

BART [27] és una arquitectura que es basa en els *transformers sequence-to-sequence*[2] estàndard amb codificació bidireccional sobre documents. Per a l'entrenament de models, els documents experimenten una sèrie de transformacions que el sistema ha d'aprendre a resoldre o corregir pel sistema de descodificació per tal obtindre el text objectiu. [3]. D'aquesta manera els models aconsegueixen complir tasques de diferents branques del PLN com la traducció i generació de textos o la classificació de documents.

[2] Podeu consultar l'article [31] per saber més d'aquest sistema.

[3] Algunes d'aquestes modificacions consisteixen en la supressió d'elements, emmascarament de paraules, permutació de paraules i oracions i rotació de documents, On 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.

Principalment BART destaca en les seues prestacions per a tasques de generació de text, tot i això, els seus resultats per a la compressió de documents també obtenen bons resultats. Per a tasques com compressió de diàlegs abstractius, cerca de respostes i generacions presenta resultats que serveixen de referent i amb resultats propers als millors sistemes.

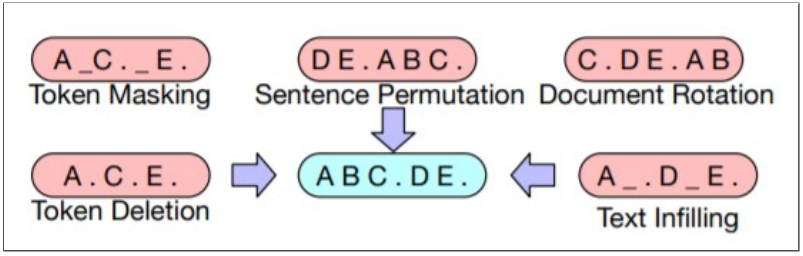

Figura 3.6: Visualització de l'emmascarament i soroll aplicat en l'entrenament dels models de BART [27].

a partir de l'entrenament d'un auto-encoder basat en BART es va crear mBART [28]. Aquest sistema, dissenyat originalment per a la realització de tasques de traducció, ha sigut preentrenat amb conjunts de documents monolingües. En aquest treball es farà servir el model de mBART entrenat pel grup ELiRF per a l'obtenció de resums a partir del corpus DACSA per tal d'obtindre resums abstractius de les notícies i comparar les prestacions dels textos processats.

3.6.2. mT5

D'igual manera que mBART, mT5 [30] es pot definir com la versió multilingüe massiu de T5 [29], un sistema destinat a realitzar tasques de PLN conegut per les bones prestacions dels models en diverses disciplines. Per la seua importància en l'entorn del NPL també utilitzarem els resums que genere mT5 en català per a l'avaluació del treball realitzat al llarg del present treball.

3.6.3. NASCA

NASCA [16], *News Abstractive Summarization for Catalan*, és un model de encoder-decoder amb *Transformers* amb la mateixa arquitectura que BART. Amb l'objectiu d'incrementar el coneixement sobre el llenguatge del model i l'abstractivitat dels seus resultats, combina múltiples tasques de preentrenament amb documents de múltiples corpus en català.

Aquest model, tractant-se d'un model monolingüe aconsegueix resultats per a la generació automàtica de resums resultats molt similars als models multilingües punters i fins i tot els supera en algunes de les mètriques i en mesures d'abstractivitat.

L'ús d'aquest tipus de sistema és de gran rellevància, ja que mostra com models entrenats per a una única llengua i amb una quantitat molt menor de dades d'entrenament es poden aconseguir resultats que competeixen amb els millors models actuals. A més, també resulta intuïtiu fer ús de sistemes especialitzats en la llengua catalana per tal de generar els resums i comparar els resultats del preprocessat de les notícies en català.

CAPÍTOL 4
Ferramentes utilitzades

En aquest capítol exposarem quines han sigut les diferents ferramentes, tant de software com hardware, que s'han fet servir per al tractament de les dades i la realització de l'experimentació i la importància de la seua aportació a les diferents tasques realitzades.

4.1 Entorn Software

A continuació presentem els principals recursos software que s'han fet servir per a la realització del treball i l'experimentació i la seua rellevància en les diferents tasques.

4.1.1. Llenguatge de programació

Com a principal recurs de Software que s'ha utilitzat en el treball trobem el llenguatge de programació Python. Python és un llenguatge de programació d'alt nivell, interpretat i de caràcter general. La seua filosofia es centra en la fàcil legibilitat del codi. Aquesta qualitat el fa atractiu per a començar a programar i fàcil d'aprendre fomentant la publicació de codis en aquest llenguatge en diferents àmbits. A més, gràcies al fet que ha guanyat popularitat el seu ús per a la realització de scripts, trobar programes ràpids d'utilitzar i de molt alt nivell resulta molt còmode. Aquestes qualitats fan que gaudisca d'una gran popularitat tot i que, de manera general, el seu codi no gaudeix de la mateixa eficiència que altres llenguatges.

Amb aquesta popularitat, el desenvolupament de ferramentes d'accés lliure no para d'augmentar, facilita la realització de programes en molts àmbits diferents. És per aquest motiu pel qual hem decidit fer servir Python en el nostre projecte. Com exposarem més endavant, Python presenta un gran nombre de llibreries ben documentades i fàcils d'utilitzar destinades al tractament de dades, aportació d'utilitats o al desenvolupament i utilització de sistemes basats en xarxes neuronals.

A més a més, com s'ha vist en el transcurs de diferents assignatures del màster cursat, Python és un dels llenguatges de programació més populars en l'àmbit de la recerca, junt amb R. Per aquest motiu és fàcil trobar papers i articles que aporten codi en Python que faciliten la utilització, adaptació i enteniment dels experiments realitzats en la recerca. La situació és la mateixa en l'àmbit de l'aprenentatge automàtic i el PLN, trobant accessibles per a Python versions de models, sistemes i corpus destinades a les principals branques de recerca de les diferents disciplines.

4.1.2. Llibreries

En la realització del nostre treball s'ha fet ús de llibreries com jsonlines, csv, statistics i numpy per a tasques habituals com la manipulació de documents i tractament d'estructures de dades. Per a l'elaboració i utilització de sistemes com NASca i BART o l'entrenament i avaluació de models s'ha fet ús d'altres llibreries més complexes que expliquem a continuació.

NLTK

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

Keras

Keras és una llibreria de Python de codi obert que facilita la utilització de TensorFlow, una coneguda llibreria que facilita el tractament i creació de xarxes neuronals. Keras proporciona una sèrie d'utilitats que permeten desenvolupar fàcilment aplicacions relacionades amb l'aprenentatge automàtic. A més permet la realització eficient d'operacions, especialment els relacionats amb la computació numèrica, i permet crear ràpidament models d'aprenentatge profund i entrenar-los i utilitzar-los amb gran facilitat. La utilització de Keras en aquest treball la trobem en la configuració, entrenament i aplicació del model de SHANN i reentrenament del model de NASca.

Datasets

Datasets es tracta d'una llibreria gestionada per l'equip de HuggingFace i creada a partir de TensorFlow datasets. Les seues principals funcionalitats són la ràpida càrrega de *datasets* i el preprocessat eficient de dades. També proporciona accés a una gran quantitat de mètriques i altres utilitats per al tractament de documents [32]. En aquest treball hem fet servir la llibreria per a l'aplicació de les mètriques de Rouge i BertScore.

Transformers

D'igual manera que Datasets, Transformers es tracta d'una llibreria de l'equip de HuggingFace. Aquesta llibreria proporciona una gran quantitat de models entrenats per a tasques de visió, àudio i text. A més permet la ràpida descàrrega i utilització d'aquests models i és compatible amb les principals llibreries destinades a l'Aprenentatge Profund (com són Jax, Pytorch i Tensorflow). En el present treball s'ha fet ús d'aquesta llibreria per a obtindre models preentrenats de mBART i mT5 reentrenats per a la tasca de generació automàtica de resums per a textos del corpus DACSA.

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. 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. Els *embeddings* utilitzats en el projecte es caracteritzen per limitar a 100 el nombre mínim d'aparicions d'una paraula per a incloure en el model la seua representació vectorial i no hi ha una diferenciació entre majúscules i minúscules.

4.2 Entorn Hardware

En el present treball no s'ha realitzat cap estudi ni aportació relacionada amb la importància del Hardware en l'execució de models de generació automàtica de resums. Tot i això, com en la resta de tasques on es fa ús de sistemes basats en xarxes neuronals, cal esmentar l'ajuda que suposa disposar de targetes gràfiques. Tant en els models utilitzats per a l'experimentació com en altres ferramentes utilitzades en aquest projecte les operacions matricials són predominants i suposen una gran consumició de temps que es veu alleugerada en part pel funcionament de les targetes gràfiques.

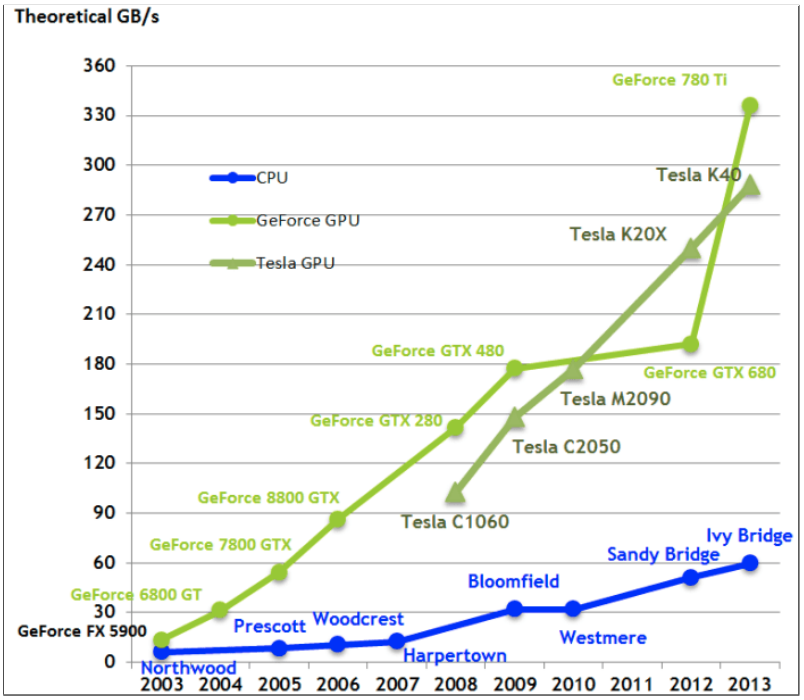

Figura 4.1: En aquesta gràfica es mostra la importància que pot aplegar a tindre l'ús de targetes gràfiques quant a eficiència temporal.

En l'actualitat els treballs basats en xarxes neuronals cada vegada presenten sistemes més grans, augmentant també així la dependència de les targetes gràfiques i la seua eficiència. Amb aquest fenomen es fa palpés la importància de disposar de recursos de grups de recerca o institucions científiques per poder aspirar a tindre importància en el món de la recerca seguint la tendència general. També cal destacar que cada vegada és més popular recórrer a plataformes de processament remotes i de pagament proporcionades per grans entitats tecnològiques com són Google o Microsoft.

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 d'una manera senzilla. En aquest projecte s'ha utilitzat dues targetes gràfiques principal-

ment, una NVIDIA Titan X, proporcionada per la màquina TARDIS del grup de recerca on s'ha realitzat el treball, i una NVIDIA GEFORCE GTX 1650 d'àmbit particular.

CAPÍTOL 5
Experimentació i resultats

En aquest capítol expliquem els diferents estudis i experiments realitzats per a l'elaboració del present treball. En primer lloc presentarem les diferents mesures que s'han realitzat sobre el Corpus i les decisions considerades en conseqüència. A continuació explicarem el preprocessat de les notícies mitjançant mètodes extractius. Seguidament presentarem l'obtenció dels resums mitjançant sistemes de generació de resums abstractius i l'aplicació de les diferents mètriques sobre els resultats. Finalment valorarem els valors obtinguts i avaluarem el procés seguit a l'experimentació.

5.1 Anàlisi i preprocessat del corpus

5.1.1. Característiques del corpus

Per poder realitzar una correcta experimentació i avaluació dels resultats és necessari conèixer les característiques de les dades que treballem. Com hem comentat anteriorment, les notícies tendeixen a presentar la informació més important a les frases inicials, per aquest motiu els sistemes que es centren en aquestes dades solen presentar bons resultats.

En el cas de NASCA, el sistema monolingüe de generació de resums abstractius que farem servir a l'experimentació, trobem al seu funcionament que el document rebut a l'entrada es dividit en subwords[1] i només es té en consideració les primeres 512 subwords de la notícia.

Per a fer un enfocament adequat del treball hem trobat necessari saber quin volum de notícies es veu completament considerat en aquesta limitació i quin percentatge de text és ignorat. En la gràfica de 5.1 podem observar que les notícies, en la gran majoria, superen el nombre de *subwords* seleccionades pel sistema NASCA. Comptant les frases que queden fora del límit de 512 *subwords* observem que més del 52% de la informació present al corpus queda exclosa en l'entrada. Per poder augmentar aquest valor s'ha contemplant l'opció d'augmentar l'entrada de 512 a 1024, valor que utilitzen sistemes com BART. Tot i això, si s'adaptaren els models i tractaren d'igual manera les notícies amb una entrada del doble de grandària, la informació que es tindria en consideració augmentaria al 87%.

A partir d'aquesta informació es poden seguir dos camins alternatius amb la finalitat d'obtindre millors prestacions, millorar la qualitat de la informació proporcionada al

[1] Les subwords són divisions del text en una mesura diferent del token habitual, normalment de mida menor al token. Tot i això, en alguns casos pot incloure fragments de diferents paraules. En aquest sistema els *subwords* extrets venen determinats per un tokenitzador que proporciona el mateix sistema NASCA.

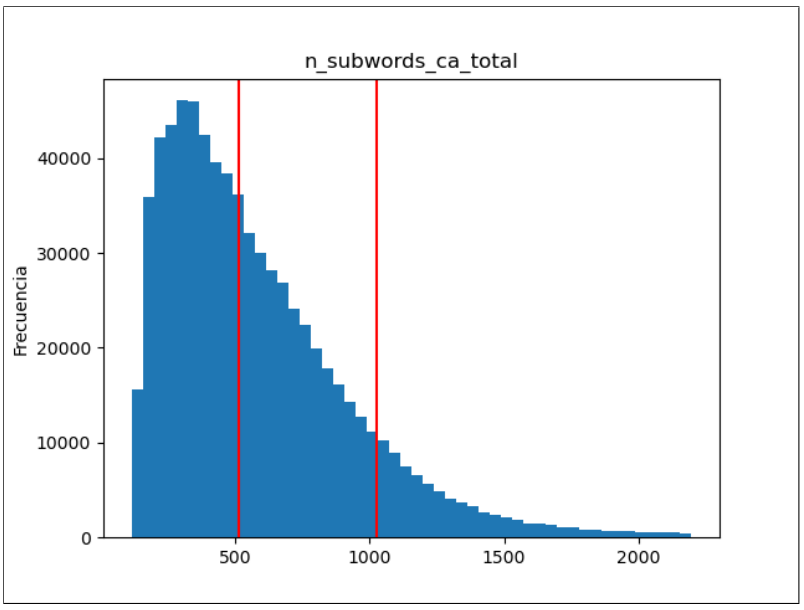

Figura 5.1: Mostra la distribució de longituds de les notícies i destaca el punt on es superen les 512 i les 1024 subwords.

sistema o augmentar la quantitat de *subwords* proporcionades a l'entrada del model. En aquest treball hem pres la decisió d'adoptar la primera opció i així evitar els inconvenients de tractar d'optar cada vegada per models més grans.

Com hem comentat anteriorment, les notícies presenten habitualment la informació rellevant al principi d'aquesta. En conseqüència, hem decidit obtindre els valors de ROUGE de les diferents particions per conèixer el grau d'importància d'aquest fenomen en el corpus que treballem. Els resultats de ROUGE han sigut multiplicats per 100 per poder realitzar una millor comparació i s'hi han obtingut considerant els resums de referència de les notícies i les dues primeres frases d'aquestes.

Font Periodística	**Test**$_V$			**Test**$_{NV}$		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Diari ARA	22,20	08,80	16,82			
Diari de Griona	25,14	10,35	18,55			
Diari La Veu	22,07	09,96	16,82			
El Per. de Catalunya	24,01	08,99	17,16			
El Punt Avui				27,04	10,77	19,14
El Temps				25,87	07,00	15,91
Nació Digital	22,26	07,16	15,80			
Regió 7	24,83	10,69	18,53			
VilaWeb				44,63	33,61	39,56

Taula 5.1: Resultats de l'anàlisi de puntuació Lead-2 per fonts per a les particions de $test_V$ i $test_{NV}$

A les taules 5.1 i 5.2 podem observar que a totes les particions es presenta un valor mitjà molt similar, reafirmant així l'adequada distribució de les notícies a les particions.

Quant a les fonts, totes presenten puntuacions similars, les quals es poden considerar acceptables. Els valors al voltant de 20 expressen que, de manera general, el començament de les notícies presenta una important relació amb el resum de referència, però no presenta un valor excessivament elevat de similitud. El valor més destacat el trobem a la font de *VilaWeb* amb una puntuació de quasi 45. Tot i això, al tractar-se d'una font exclusiva de la partició de $test_{NV}$ no tindrà cap impacte a l'entrenament de models.

Font Periodística	**Training**			**Validation**		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Diari ARA	22,06	08,79	16,80	22,01	08,75	16,78
Diari de Griona	25,32	10,45	18,60	25,44	10,62	18,71
Diari La Veu	22,03	09,91	16,75	22,14	10,13	16,85
El Per. de Catalunya	23,80	08,79	16,97	23,82	08,82	17,06
El Punt Avui						
El Temps						
Nació Digital	22,41	07,14	15,78	22,35	07,13	15,78
Regió 7	24,67	10,58	18,55	24,58	10,43	18,46
VilaWeb						

Taula 5.2: Resultats de l'anàlisi de puntuació Lead-2 per fonts per a les particions de *training* i *validation*.

5.1.2. Preprocessat amb SHANN

Per a la millora dels textos proporcionats als diferents sistemes abstractius s'ha decidit fer ús de SHANN. En el seu funcionament SHANN atorga a cada frase del text una puntuació i fa servir les millors per a posteriorment generar el resum del text. En el present treball hem treballat amb una versió modificada de SHANN per poder obtindre la llista de frases puntuades amb un determinat format en lloc del resum de la notícia. Una vegada obtingudes les frases puntuades s'ha seleccionat les millors mantenint l'ordre original d'aparició a la notícia fins a completar el límit de 512 *subwords* establit pel model de NASCA.

Per a l'obtenció del corpus preprocessat ha sigut necessari fer ús d'embeddings en català per a l'entrenament de SHANN. Els embeddings que s'han fet servir han sigut entrenats amb els documents en català del corpus OSCAR, la Viquipèdia i els textos descartats del corpus DACSA. Com a paràmetres de l'entrenament es va prendre un mínim de 10 aparicions per paraula per a mantindre l'*embedding* i distinció entre majúscules i minúscules. L'entrenament del mateix sistema SHANN s'ha realitzat fent ús dels conjunts de training i validació del Corpus DACSA. En el quadre 5.3 observem les característiques dels *embeddings* utilitzats, el percentatge de tokens cobert i el percentatge de vocabulari[2] que es veuen representats pels vectors.

Partició	**Tokens**	**Vocabulari**
Test_V	99,36%	69,92%
Test_{NV}	99,33%	68,92%
Training	99,43%	30,72%
Validation	99,39%	70,56%

Taula 5.3: 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*.

Aquests *embeddings* han sigut obtinguts a nivell de paraula. Com que el conjunt de *training* ha sigut utilitzat per a l'entrenament podem veure com presenta un cobriment del total de tokens (paraules, en aquest cas) lleugerament superior al de les altres particions. També s'observa una gran diferència entre el percentatge de paraules representades i el percentatge de vocabulari. El principal causant d'aquest fenomen és el requisit dels tokens d'apareixer 10 vegades per a tindre un vector que el represente als *embeddings*.

[2]Definim com *Vocabulari* al conjunt de tokens o paraules úniques que podem trobar al conjunt de notícies del corpus.

Moltes aparicions d'emoticones, errors ortogràfics o noms propis són molt concrets o molt poc freqüents i no es consideren de gran rellevància per a ser representats als *embeddings*.

En el cas del conjunt d'entrenament veiem que es trenca la tendència marcada als altres conjunts. Açò es deu al fet que, tot i que s'ha entrenat els *embeddings* amb aquest conjunt, al presentar un nombre de notícies molt major (de l'ordre de 18 vegades més que la resta) el percentatge de vocabulari que no es repeteix més de 10 vegades és molt més elevat. En aquestes taules veiem com al considerar paraules completes el vocabulari pot ser molt distint i, a més de requerir un major emmagatzematge, molta informació no s'aplega a contemplar i es perd per a la realització de les diferents tasques a abordar. Aquest inconvenient s'ha alleugerat recentment en els sistemes abstractius de generació automàtica de resums al ser capaços d'utilitzar fàcilment estructures diferents dels tokens com són els *subwords*, ja que no necessiten extraure fragments directament del text com els sistemes extractius.

5.2 Obtenció de resums

D'acord amb l'anàlisi anterior s'ha realitzat l'obtenció dels resums per diferents sistemes abstractius de generació automàtica de resums. Els resultats s'hi han obtingut tant per a test_V com per a test_{NV} per a tots els sistemes. Per a tots els sistemes s'han generat els resums tant aplicant el preprocessat de SHANN com per al corpus original. A continuació explicarem els processos d'obtenció portats a terme i els resultats de les mètriques sobre els resums generats.

5.2.1. NASCA

Per al cas de NASCA s'ha fet servir la versió oficial de Huggingface del sistema i els models que proporciona. Com que els models proporcionats per aquest sistema estan ja entrenats amb el corpus DACSA no ha sigut necessari realitzar ningun tipus d'entrenament del model per a adaptar-lo al corpus. Tot i això sí que s'ha realitzat un procés de fine-tunning amb el conjunt d'entrenament per a comparar les prestacions del model després de presentar la nova estructura de les notícies en fase d'entrenament.

Partició	Rouge-1	Rouge-2	Rouge-L	BERTScore
Test_V	28,74	11,58	22,68	71,77
Test_{NV}	28,22	11,27	21,49	70,18

Taula 5.4: Model base de NASCA amb corpus original.

Partició	Rouge-1	Rouge-2	Rouge-L	BERTScore
Test_V	27,58	10,66	21,70	71,31
Test_{NV}	28,22	11,30	21,50	70,18

Taula 5.5: Model base de NASCA amb corpus preprocessat.

Partició	Rouge-1	Rouge-2	Rouge-L	BERTScore
Test_V	28,66	11,46	22,51	71,61
Test_{NV}	27,90	11,09	21,21	70,02

Taula 5.6: Model reentrenat de NASCA amb corpus preprocessat.

En els resultats de 5.4 i 5.5 podem veure com sense cap tipus de reentrenament el sistema NASCA es veu afectat negativament en totes les mètriques per al conjunt de Test_V en major o menor mesura. En el cas del Test_{NV} en canvi trobem que els resultats milloren lleugerament. Açò es pot deure a dos motius principals, el primer és que el preprocessament amb SHANN aproxima l'estructura de les fonts no vistes a les del conjunt d'entrenament, ajudant així a la selecció del resum per part de NASCA. Una altra explicació la dona el fet que fonts com *El Temps*, en ocasions presenta notícies molt llargues, podent quedar fora de l'entrada de NASCA parts rellevants. Aquesta última opció explicaria millor la lleugeresa del canvi en els resultats.

Una vegada s'ha portat a terme el reentrenament podem veure en 5.6 com els resultats de Test_V milloren significativament respecte a la taula anterior 5.5 però segueixen sent menors que els valors originals. Açò porta a pensar que aquesta estratègia efectivament modifica la manera en la qual s'ha de prestar atenció als elements de l'entrada per a generar els resums. Com que el reentrenament no s'ha portat fins a convergència a causa de les limitacions de temps i recursos no podem determinar com de proper estan els resultats obtinguts del màxim raonable.

Per altra banda, veiem com en el cas del Test_{NV} pasa el contrari que en el cas anterior, els resultats són lleugerament inferior, tant dels valors originals com en el model sense reentrenar però amb preprocessat. En aquesta ocasió el sistema s'ha acostumat a tractar els documents preprocessats del conjunt de Test_V, cosa que aparentment perjudica els resultats de notícies d'altres fonts.

5.2.2. mBART

Per a la utilització de mBART amb el corpus de DACSA s'ha decidit fer ús del model ja entrenat pel propi grup de recerca ELiRF *mbart-large-cc25-dacsa-ca*[3]. Aquest model està generat a partir del model multilingüe *mbart-large-cc25* publicat online per Facebook[4] i ha sigut entrenat per a la tasca de generació automàtica de resums a partir del corpus DACSA.

En el cas de mBART veiem com els resultats abans, 5.7, i després, 5.8, del preprocessat del corpus els valors de les mètriques utilitzades no presenten canvis significatius a les mètriques, tant en el conjunt de Test_V com en el de Test_{NV}.

Partició	Rouge-1	Rouge-2	Rouge-L	BERTScore
Test_V	27,60	11,43	22,11	67,04
Test_{NV}	27,15	10,97	20,81	66,26

Taula 5.7: Resultats de mBART amb el corpus original.

Partició	Rouge-1	Rouge-2	Rouge-L	BERTScore
Test_V	27,60	11,43	22,11	67,04
Test_{NV}	27,15	10,97	20,80	66,26

Taula 5.8: Resultats de mBART amb el corpus preprocessat.

[3]https://huggingface.co/ELiRF/mbart-large-cc25-dacsa-ca
[4]https://huggingface.co/facebook/mbart-large-cc25

5.2.3. mT5

D'igual manera que amb el sistema anterior, amb mT5 hem fet servir un model entrenat pel grup en el qual hem treballat, *mt5-base-dacsa-ca*[5]. En aquest cas el model original[6] va ser publicat per Google i també ha sigut entrenat únicament per la generació de resums a partir de les notícies del corpus DACSA.

D'igual manera que passa amb els models de mBART, el preprocessament generat per SHANN no pareix modificar els resultats de ROUGE i BertScore per als diferents conjunts d'entrenament.

Partició	Rouge-1	Rouge-2	Rouge-L	BERTScore
Test_V	25,95	10,30	20,95	66,56
Test_{NV}	26,01	10,83	20,46	66,11

Taula 5.9: Resultats de mT5 amb el corpus original.

Partició	Rouge-1	Rouge-2	Rouge-L	BERTScore
Test_V	25,95	10,29	20,94	66,56
Test_{NV}	26,01	10,83	20,46	66,11

Taula 5.10: Resultats de mT5 amb el corpus preprocessat.

5.3 Avaluació dels resultats

A continuació compararem els diferents resultats dels experiments realitzats i comentarem les diferents observacions que es poden fer sobre els resultats obtinguts. Comentarem els resultats presentats a les taules dels diferents sistemes de generació de resums i comentarem les diferències que es poden observar entre els resums obtinguts segons les dades de test utilitzades.

Quant als resultats dels resums generats pels models de NASCA observem que en el model sense reentrenar l'aplicació del corpus preprocessat ha tingut dos efectes distints. En el conjunt de Test_V, que és el que presenta notícies de les mateixes fonts amb les quals s'ha entrenat el model, els resultats de les diferents mètriques s'han vist lleugerament afectades de manera negativa. Tot i això, amb aquest mateix model i seguint el mateix procés de preprocessat el conjunt de Test_{NV}, el qual presenta notícies no vistes a l'entrenament, ha vist els seus resultats millorats després d'aplicar l'entrenament.

A continuació veiem com al reentrenar el model de NASCA amb el corpus preprocessat amb SHANN els resultats canvien de tendència. Els resultats de les fonts vistes es veuen notòriament afavorits tot i que el reentrenament no va poder fer-se fins la convergència del model. Aquesta millora no aconsegueix superar els resultats del model de NASCA original però sí que modifiquen en alguns casos les frases en les quals es basa el sistema per a generar els resums, com podem observar a continuació.

Notícia

[5] https://huggingface.co/ELiRF/mt5-base-dacsa-ca
[6] https://huggingface.co/google/mt5-base

Una vintena d'organitzacions de professionals de la sanitat, pacients i consumidors han demanat avui al Senat que s'impedeixi qualsevol excepció que possibiliti fumar en els llocs d'hostaleria, oci i restauració a la reforma de la llei antitabac. La Comissió de Sanitat de la Cambra Alta debat avui les esmenes presentades a la reforma de la llei antitabac (2005) perquè no es pugui fumar en cap espai públic tancat, inclosos els d'hostaleria, oci i restauració, els únics on fins ara podia fer-se amb certes condicions. "Els 'cubicles' serien l'excusa perfecta per no complir la llei i mantenir l'omnipresència del fum en els espais d'oci i convivència", segons adverteix una carta oberta d'aquestes organitzacions dirigida als representants polítics en relació amb les excepcions i moratòries que introdueixen algunes esmenes. El president de l'Organització Mèdica Col.legial (OMC), Juan Antonio Rodríguez Sendín, s'ha preguntat en la presentació de la carta "quin és el preu"de cada mort evitable per fum de tabac. La presidenta del Comitè Nacional de Prevenció del Tabaquisme (CNPT), Ángeles Planchuelo, ha subratllat que les excepcions no protegeixen els treballadors ni als usuaris. "Generen -ha prosseguit- un pèssim model educatiu per als menors d'edat i contribueixen a mantenir la normalitat social del consum de tabac, aspecte aquest últim de gran interès per a la indústria del tabac a l'hora de mantenir el seu nivell de negoci". El document, presentat amb l'assistència de representis de les organitzacions, assegura que "no té cap justificació"que es busquin fórmules per seguir contaminant els espais públics tancats per "raons econòmiques que no tenen fonamentï és ïntolerable"que es puguin defensar per interessos d'oportunitat política. "Rebutgem enèrgicament -diu la carta- que tots dos interessos s'anteposin a l'obligació de preservar la salut dels nostres ciutadans". Recorden les organitzacions que el fum del tabac mata cada any a Espanya a 56.000 persones, de les quals 3.000 moren per respirar en contra de la seva voluntat l'aire contaminat pel fum del tabac.

Resum de referència

Una vintena d'organitzacions de professionals de la salut, pacients i consumidors han demanat aquest dimecres al Senat que s'eviti qualsevol excepció en la nova llei.

Resum base

La Comissió de Sanitat de la Cambra Alta debat avui les esmenes presentades a la reforma de la llei antitabac perquè no es pugui fumar en cap espai públic tancat.",

Resum reentrenat i preprocessat

Una vintena d'organitzacions de professionals de la sanitat, pacients i consumidors demanen al Senat que s'impedeixi fumar en els llocs d'hostaleria, oci i restauració.",

En aquest exemple veiem com amb el reentrenament i amb el nou corpus el model és capaç d'obtindre un millor resultat que en el cas base. Tot i que en la notícia original el resum adequat es troba a les primeres frases de la notícia el model de NASCA no troba que aquesta informació siga rellevant fins a haver realitzat el reentrenament amb el corpus del conjunt d'entrenament processat per SHANN.

Notícia

La plaça d'Espanya de Barcelona serà el centre neuràlgic de la manifestació independentista de l'Onze de Setembre d'enguany. La mobilització, organitzada per l'ANC i amb el suport d'Òmnium Cultural i l'AMI, dibuixarà un estel que ocuparà els carrers Tarragona, Creu Coberta i Gran Via i les avingudes del Paral·lel i de Maria Cristina, amb la plaça com a punt de confluència. "Volem representar que des de camins diferents hi ha punts de trobada on ens posem d'acord per un objectiu comú", va explicar la presidenta de l'ANC, Elisenda Paluzie, en la presentació de la mobilització, que enguany té per lema "Objectiu independència". Els organitzadors fan una crida a la participació perquè insisteixen: "Quan la societat civil mobilitzada, el Govern i la majoria independentista al Parlament caminem plegats, som capaços d'arribar molt lluny". Per això demanen que la ciutadania es mobilitzi i "ompli els carrers". Alhora, demanen que la gent que vulgui participar-hi s'hi inscrigui perquè tots els trams quedin plens i l'acte pugui transcórrer ordenadament. Çal tornar a posar la independència al centre del debat, només així podrem afrontar els reptes que s'albiren a l'horitzó", insisteixen. Com és habitual en les mobilitzacions de la Diada, la manifestació es divideix en trams per assegurar una distribució homogènia dels participants. Aquest cop n'hi ha 26 i, de moment, ja hi ha 400.000 persones inscrites, 250.000 samarretes venudes i 1.200 autocars, segons l'ANC. L'ocupació dels trams és, ara mateix, la següent. Pots inscriure't a la mobilització de la Diada aqu. Les inscripcions es poden fer a través de la pàgina web o per telèfon, i de manera individual o en grup. La Diada per la República començarà a les 17.14 hores, però l'organització recomana ser al tram assignat a les 16 h.

Resum de referència

Tota la informació sobre la mobilització de l'Onze de Setembre d'enguany organitzada per l'ANC i Òmnium Cultural.

Resum base

La plaça d'Espanya de Barcelona serà el centre neuràlgic de la mobilització independentista de l'Onze de Setembre.

Resum reentrenat i preprocessat

La mobilització, organitzada per l'ANC i amb el suport d'Òmnium Cultural i l'AMI, dibuixarà un estel que ocuparà els carrers Tarragona, Creu Coberta i Gran Via.

En aquest cas els resultats de les mètriques dels dos valors generats són molt similars, tot i això els resums són totalment distints. Ambdós resums presenten elements que es poden trobar al resum de referència encara que cap dels dos descriga exactament la mateixa informació. En aquest cas queda visible que la manera de generar els resums amb el corpus original i el preprocessat es veu afectada de maneres que no es poden visualitzar clarament amb mètriques reconegudes com són ROUGE o BertScore.

En el cas de mBART hem vist que els resultats per als resums dels dos conjunts de testeig no s'han vist significativament modificats pel preprocessat amb el sistema extractiu. Els mateixos resultats els podem observar en les mètriques dels resums generats pels models de mT5. Aquests resultats, tot i que a les mètriques no reflecteixen cap canvi significatiu, han sigut obtinguts generalment amb una entrada de menor mida que la que obtenen amb el corpus original.

Com hem vist a les taules 5.1 i 5.2 les notícies del corpus obtenen uns bons resultats en les mètriques de ROUGE en sistemes tradicionals com Lead-2. Com hem comentat

anteriorment aquests valors es deuen a l'estructura habitual de les notícies. Tenint en compte aquest fet s'entén que la influència del preprocessat realitzat no haja mostrat un impacte palpable als valors de les mètriques.

En aquesta valoració hem vist com el preprocessat dels textos ha aconseguit que amb una entrada reduïda els models aconseguisquen obtindre els mateixos resultats en la gran majoria de casos, com és el cas dels models de mT5 i mBART. També hem vist que en alguns casos, per a documents de fonts no vistes en l'entrenament aconsegueix presentar millors resultats, com és el cas de NASCA. Per últim, hem vist en el reentrenament de NASCA que, amb l'adequat temps de reentrenament dels models, aquest procés presenta el potencial per a obtindre millors resultats que en les models tradicionals, sobretot en el cas de documents amb menor presència d'informació rellevant a les primeres frases del text.

CAPÍTOL 6
Conclusions

En aquesta memòria hem vist les diferents decisions preses al llarg del treball, les decisions preses per al procés de selecció de contingut del corpus, 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. Aquestes dades, a més de per conèixer millor el corpus, ens ha ajudat directament en la presa de decisions a l'hora de realitzar les diferents tasques del projecte. D'aquesta manera podem considerar que l'objectiu del treball **Anàlisi de les característiques pròpies del corpus** es considera completat. El segon objectiu, **Processament de textos de notícies mitjançant sistemes de resum extractiu**, també s'ha complit totalment, ja que s'ha entrenat el sistema extractiu SHANN per a la selecció del contingut que s'utilitzarà per a l'elaboració dels resums a estudiar.

Una vegada obtingut el corpus preprocessat s'ha entrenat i fer servir models de NASCA, mBAR i mT5 per a la generació automàtica de resums abstractius. A més, cada un d'aquests models s'ha fet servir en diverses ocasions per tal de poder comprovar els efectes de l'aplicació prèvia de SHANN sobre les notícies de DACSA. Com que els tres sistemes emprats són sistemes basats en xarxes neuronals podem concloure que el tercer objectiu, **Utilització de sistemes basats en xarxes neuronals per a la generació de resums**, també s'ha completat. L'objectiu d'**avaluació dels 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.

Com s'ha comentat al llarg del treball el propòsit general d'aquest projecte era l'avaluació de la influència de la selecció de contingut per a l'entrada d'un sistema de generació automàtica de resums. Amb aquest objectiu calia esperar que la selecció de frases per un sistema extractiu de bones prestacions com és SHANN millorara els resultats dels models entrenats amb el corpus original. Tot i que aquest resultat no s'ha aplegat a manifestar sí que hem pogut veure que el procés d'aplicació de SHANN sí que ha modificat la manera d'obtindre els resums en un nombre significatiu de notícies. El comportament dels sistemes porta a pensar que l'estructura de les notícies, les quals presenten la informació més rellevant a l'inici del text, suposa una influència major als resultats que la que pot aplegar a realitzar la selecció de contingut estudiada en el present treball.

Per últim, hem vist en els resultats de la nostra experimentació que en els sistemes com mBART i mT5 l'entrada de menor grandària generada a l'aplicar SHANN ha sigut capaç d'obtindre, en mitjana, resultats pràcticament iguals als obtinguts amb els corpus original i de major mida. D'aquesta manera veiem que la reducció de la dimensionalitat

de l'entrada té potencial per obtindre els mateixos resultats amb un cost menor en els sistemes de generació automàtica de resums.

6.1 Treball futur

Com hem comentat al llarg de la present memòria, aquest treball podia seguir diferents línies de treball partint de diferents consideracions per tal de portar a terme l'objectiu d'aquest treball. Aquestes són algunes tasques que es podrien realitzar basant-nos en els resultats del nostre projecte o com a alternativa per a ampliar l'experimentació ja realitzada:

- **Reduir l'entrada dels models**: En aquest treball hem observat com la selecció del contingut per a l'entrada de sistemes de generació automàtica de resums és una manera eficaç de reduir la dimensió de l'entrada sense afectar negativament als resultats. Coneixent aquesta informació, un treball interessant a realitzar seria l'elaboració de sistemes que disminuïsquen de manera important la mida de l'entrada reduint la grandària dels models i aplicar preprocessaments a l'entrada per a la selecció del contingut amb la informació de major rellevància. D'aquesta manera es desenvoluparia en major profunditat l'objectiu d'aconseguir obtindre models amb menors requisits temporals i espacials i que obtinguen resums de les mateixes característiques que sistemes de major envergadura.
- **Treballar amb altres tipus de text**: Com hem comentat anteriorment, els articles periodístics presenten una estructura on la creació del resum a partir de les primeres frases es veu afavorida. Tot i això, el present treball ha sigut capaç de mantindre els resultats després del preprocessat o fins i tot millorar-los lleugerament en el cas de fonts no vistes. En documents on la informació principal no es trobe al començament del document cal esperar que els sistemes de generació de resums es vegen perjudicats al presentar una limitació de mida a l'entrada. Amb el procés de selecció de contingut com el que hem fet servir en el present treball seria raonable esperar que la informació rellevant que reben els sistemes augmentara de manera important i fora determinant per a la correcta generació del resum. Un altre tipus de resum que es podria veure afavorit per aquest procés seria el resum de múltiples documents. D'aquesta manera el sistema destinat a la generació del resum podria rebre les frases més importants del conjunt de documents i generar el resultat de manera molt més directa amb la limitació de l'entrada que presenta en l'actualitat.

Bibliografia

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

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Vicent Ahuir, Lluís-F Hurtado, José Ángel González i Encarna Segarra NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish *Applied Sciences*, v11, 21, 9872, 2021

[17] 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.

[18] Max Grusky, Mor Naaman i Yoav Artzi NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies *Proceedings of NAACL-HLT*, 708—719, juny, 2018.

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

[20] 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.

[21] 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.

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

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

[24] Encarna Segarra, Vicent Ahuir, Lluís-F. Hurtado i José Ángel González DACSA: A large-scale Dataset for Automatic summarization of Catalan and Spanish newspaper Articles *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5931 – 5943, juliol, 2022.

[25] 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.

[26] 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.

[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] 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.

[29] 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:1910.10683*, octubre, 2019.

[30] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua i Colin Raffel mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934v3*, març, 2021.

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

[32] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush i Thomas Wolf Datasets: A Community Library for Natural Language Processing. *arXiv preprint arXiv:2109.02846*, setembre, 2021.