KONPUTAZIOA

Osasun-arloko laburduren desanbiguazioa

Gradu Amaierako Proiektua

Egilea:
Maite Laca Saralegui

Zuzendaria:
Maite Oronoz Anchordoqui

Data:
2016/06/24

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

ESKER ONAK

Eskerrik asko **Maite** proiektuko zuzendari bikaina izateagatik. Momentu zailetan beti laguntzeko prest egon zarelako eta animoak eman dizkidazulako. Eskerrik asko ere ikerketaren mundu hau ezagutzera emateagatik eta ikasitako guztiarengatik.

Eskerrik asko **Alicia** emandako laguntza guztiagatik, proposatzen zenituen ideia zein azaldu zenizkidan kontzeptu guztiengatik. Beti laguntzeko prest egon zarelako, eta beti aurrera joateko motibatzen egon zarelako.

Eskerrik asko **Aitor**, UKB-rekin izan ditudan arazo eta duda guztien aurrean, beti entzuteko eta azaltzeko prest egon zarelako.

Eskerrik asko **Sara** eta **Miriam** corpus guztia etiketatzeagatik eta proiektu honen atal garrantzitsu bat izateagatik. Pazientzia handiko lana egin duzuelako.

Eskerrik asko **unibertsitateko kide eta lagun guztioi** beti animoak emateko prest egon zaretelako, eta horrek balio handia izan duelako nigan.

Eskerrik asko **kuadrillari** beti animoak ematen egon zaretelako, gehienbat **Miriam** eta **Iratiri** azkeneko eta pauso zailenak emateko beti animoak ematen egon zaretelako.

Eskerrik asko **familia** guztiari, momentu zailetan beti laguntzeko prest egon zaretelako. Gehienbat **amari** bai momentu on eta txarretan laguntzen zaudelako eta **Leire** ahizpari beti laguntzeko prest egoteagatik.

Gaien Zerrenda

Kapitulua 1

Sarrera eta Kokapena

Medikuntzaren arloan, osasun-txostenak garrantzi handikoak dira. Osasun-txosten hauetan, informazio oso baliotsua biltegiratzen da, bai gaixoarentzako, baita medikuntzaren alorrerako era orokorrean. Osasun txostenetan, gaixoaren aurreko gaixotasunak, egin zaizkion probak, gaixotasun zehatz baterako izan dituen sintomak, edo eta, gaixotasun batentzako jaso duen tratamendua daude.

Medikuak txostenak idazten ditu, gaixoa kontsultan dituen bitartean, eta beraz, oso logikoa da, askotan testu horiek oso ondo ez ulertu edo ulergaitzak izatea, medikua gaixoari arreta jartzen dagoelako eta ez idazketan. Horrek, osasun-txosten hauek, guztiz zuzenak ez izatea ekartzen du, hau da, akats ortografiko ugari, puntuazio faltak, edo puntuazioak gaizki ipintzea, hitz bat jarri beharrean, horren akronimoak jartzea, hizkuntza ez estandarra erabiltzea, eta esaldi amaitu gabeak izatea eragiten du. Akronimo horiek dira GAP honen gai nagusia, osasun alorreko laburdurak. Laburdurak, hitzak esaten duen bezala hitz edo hedapen baten laburtzeak dira. Laburdura batek ordea, hedapen bat baino gehiago izan dezake, hau da, anbiguoak izan daitezke eta, laburduren desanbiguazioak, laguntza bat ematen du txosten hauek hobeto ulertu ahal izateko. Ikus dezagun zer den laburduren desanbiguazioa adibide batekin:

Demagun esaldi hau dugula: "ECG: RS a 95 lpm. PR normal. AQRS a -30. No bloqueos AV". Testuen ulergarritasuna handitzearren eta hauetako elementu bakoitzak zein esanahi duen jakiteko, laburdurak hedatu egiten dira. Horretarako adibidez, http://sedom.es orriko moduko laburdura medikoen hiztegiak daude. Hiztegi horretan "ECG""electrocardiograma"da baita "escala de Glasgow del coma"ere. Guk testuak analizatzen ditugunean, bi esa-

nahiak agertzen zaizkigu, hau da, laburdura anbiguoa da. Testuinguruaren arabera erabaki behar da zein den ECGrako esanahi egokia. Kasu hauetan, "lpm"ez da anbiguoa eta 'latidos por minuto' esan nahi du eta hau erabiliz, ECG 'electrocardiograma' dela esan daiteke. GAP honetan laburduren desanbiguaziorako dauden teknikak aztertu beharko dira, egokiena aukeratu eta probatu.

Proiektua, IXA taldean kokatuta dago eta talde hau, hizkuntzaren arloan sortzen diren ezaugarrien ikerketan oinarritzen da. Beraz, proiektu hau hizkuntzaren prozesamenduan oinarrituta dagoenez, hizkuntzaren ikerketan oinarritzen da. Ikerketa horretarako, IXA taldeak baliagarriak dituen txosten-medikuak erabiliko dira, baina modu murriztuan, txosten hauek konfidentzialak direlako.

Erabiliko diren osasun-txostenen hizkuntza, gaztelera izango da eta testuak Galdakao-Usansolo ospitalekoak izango dira. Hori bai, lehen esan moduan, testu hauek konfidentzialak direnez, ez dira bere osotasunean erabiliko, bere osotasunean erabiltzeko eskubiderik ez dudalako.

Horrez gain, proiektu honetan, hainbat pertsonen menpe egon beharko da: medikuak (txostenen sortzaileak), etiketatzaileak... Eta honek, hainbat oztopo izan ditzake, oso normala delako giza akatsak izatea. Arlo horretatik, adi egon beharko da.

Kapitulua 2

Proiektuaren Helburu Dokumentatua

2.1 Proiektuaren deskribapena eta helburua

Proiektu honen helburua hizkuntzaren prozesamendurako tresnek medikuntzaren arloan izan dezaketen erabilgarritasuna aztertzea da. Azterketa honetan, medikuntza arloan erabiltzen diren laburduren desanbiguazioa aztertuko da.

Medikuntzaren domeinuan, laburdura berak erabiltzen dira hainbat esanahi desberdinekin. Adibide bat emateko, *IQ* laburdurak esanahi hauek ditu:

- *Índice de Quick*
- *Intelligenz-Quotient (Coeficiente Intelectual)*
- *Intervención quirúrqica*

Eta *IQ* beharrean, *iq* bada, bere esanahia *"Alergia a las proteínas de la leche de vaca"* izango da. Beraz, testu batean, honelako terminoren bat agertzen bazaigu, pertsonok, testua interpretatuz laburduraren esanahia zein den aukeratzen dugu. Hau da, testuinguruaren arabera interpretatzen dugu zein izango den laburduraren esanahia. Baina nola egingo du hori makina batek?

Hau izango da *Gradu Amaierako Proiektu* honetan lortzen saiatuko garen helburu nagusia, hau da, makinak laburduraren hedapenen artean egokia

dena aukera dezala testuingurua kontuan izanik. Beste modu batean esanda, laburduren desanbiguazioa egitea. Horretarako, desanbiguaziorako hainbat teknika erabiliko dira, horietatik bi aukeratu eta testu errealekin ebaluazio bat egingo da, egokiena zein den aplikatzeko asmoz.

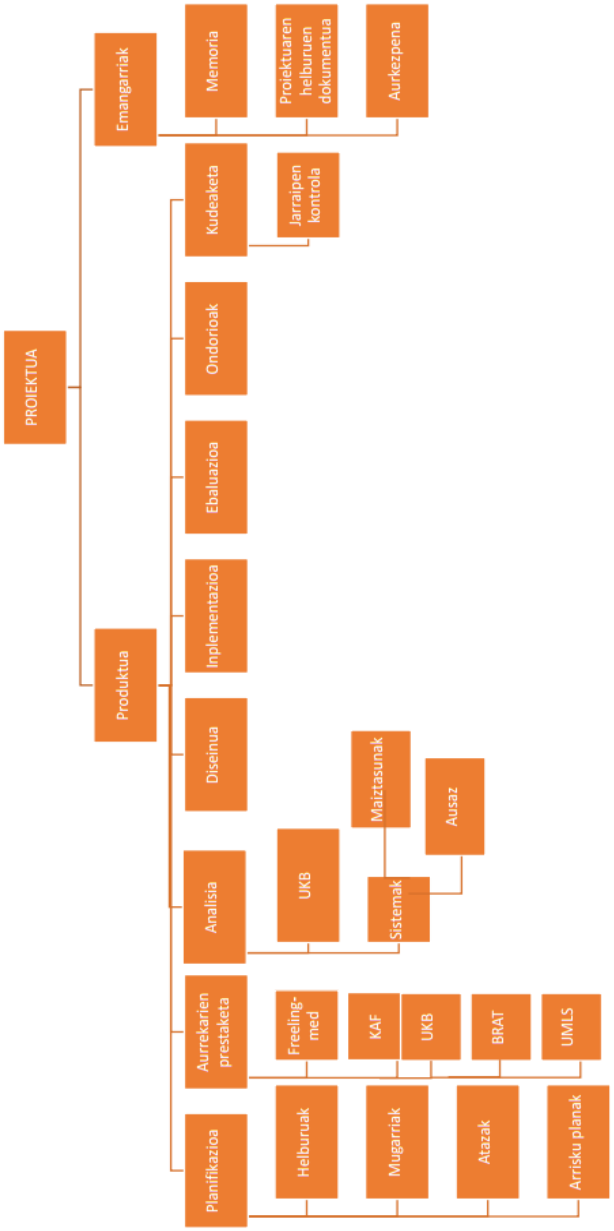
Helburu orokor honetaz gain, beste hainbat helburu baditu proiektu honek. Alde batetik, erabiliko diren fitxategi (corpus)-en pribatasunaren mantentzea helburuetako bat izango da. Hau da, erabiliko diren testu-fitxategiak, konfidentzialak izango dira eta beraz, soilik paragrafo edo esaldika hartuko dira, fitxategi osoak hartu beharrean eta modu honetan, bezeroaren konfidentzialtasuna mantenduko da.

Beste helburu bat, laburduren anbiguotasuna testu motaren arabera neurtzea izango da. Hau da, hainbat laburduren hedapena, ia beti hedapen bera izan daiteke, eta kasu horietan testu motaren arabera anbiguo edo ez-anbiguo direla esaten da. Proiektuan zehar, izatez anbiguoak diren laburdurak erabiliko dira, baina, laburdura hauek hautatutako corpusean anbiguoak diren jakitea helburu bat izango da.

2.2 Proiektuaren plangintza

2.2.1 Lanaren deskonposaketa egitura (LDE)

Jarraian, 1.1 irudian, proiektuaren lanaren deskonposaketa egitura (LDE-a) aurkezten da. Bertan proiektuan zehar burutu beharreko ataza eta azpiataza garrantzitsuenak erakusten dira.



Irudia 2.1: Lanaren deskonposaketa egitura

2.2.2 Atazen definizioa

Jarraian proiektuaren lan-pakete edo ataza ezberdinak aurkeztu eta deskribatu egiten dira:

Planifikazioa

Ataza honetan proiektua garatzeko beharrezko izango den planifikazio bat garatuko da. Bertan proiektuaren helburuak zehaztu, eta hauek betetzeko atazak definituko dira. Honez gain, atazen denbora estimazioa ere egingo da. Azkenik, proiektuaren bideragarritasun eta arrisku-plana ere garatuko dira.

Aurrekarien prestaketa

Ataza honetan proiektua garatzeko beharrezkoa den ezagutza jasoko da. Proiektua garatu ahal izateko, informazio ugari bilatu behar da. *Freeling-med* analizatzaile linguistikoa zer den eta bere ezaugarriak aztertuko dira eta *Kyoto Annotation Format (KAF)* fitxategiak nola egituratzen diren ere aztertuko da. Gainera, *KAF* fitxategietan laburdura anbiguo eta ez-anbiguoak nola agertzen diren eta zergatik, aztertuko da. Desanbiguatu beharreko laburdurak medikuntzara egokitutako *Freeling*-ek emango dizkigu eta honek irteera *eXtended Markup Languages (XML)* oinarria duen *KAF* formatuan ematen ditu.

Hauez gain, *UMLS - SNOMED CT* hiztegiei buruz ere informatu beharko da. Bi hiztegi hauek beharrezkoak izango dira, proiektuan zehar, grafoetan oinarritutako UKB izeneko tresna (ikusi 3.2.1 atala) erabiltzeko, beharrezkoak direlako, laburduren eta hauen hedapenen identifikadoreak jakitea.

Web oinarria duen *BRAT* buruz ere azalpenak emango dira. Tresna hau, GoldStandard-a sortuko duten etiketatzaileek erabiliko dute, testuko laburduren hedapenak etiketatzeko.

Azkenik, *Aurrekarien prestaketa* atalean, laburduren desanbiguazioa egiteko teknikak aztertuko dira. Hau da, ataza honetan, *Word Sense Disambiguation (WSD)*(Hitzen Adiera Desanbiguazioa) egiteko zein teknika dauden aztertuko da eta teknika horietatik probak egiteko zeintzuk erabiliko diren aukeratuko da. WSDko teknikak laburduren desanbiguaziorako erabilgarriak direla aurreikusten da.

Diseinua

Ataza honetarako helburua teknika bakoitza nola erabiliko den diseinatzea izango da. Horretarako, corpus-a nola sortu, banatu eta landuko den azalduko da.

Proiektuan zehar erabiliko diren bi sistemek behar dituzten ezaugarri guztien osaketaren azalpena egingo da atal honetan. Azalpen horien artean, *Gold Standard-aren* eraketa egongo da, adibidez.

Inplementazioa

Inplementazioan, aukeratutako teknikak aplikatzeko programak garatzen dira. Programa bakoitza zertan datzan azalduko da gainetik. Zein diren behar dituen sarrera fitxategiak eta zein ematen dituen programa bakoitzak irteera gisa.

Ebaluazioa

Ataza honetan, aukeratu diren bi sistemen emaitzak eskuz etiketatutako testu bilduma batekin konparatuko dira. Konparaketa horiek harturik, teknika onena zein den erabakiko da. Eskuz etiketatutako testu-bilduma edo corpusari *Gold Standard* deituko zaio.

Gold Standard hau, bi etiketatzaileek etiketatua izango da eta eskuz etiketatuko dituzten corpusean zehar etiketatzeko eskatuko zaizkien laburduren hedapenak.

Ondorioak

Ataza honetan proiektuari buruzko ondorio orokorrak azalduko dira. Ondorio hauen artean, bai planifikazioaren aldaketak, proiektuak prozesuan zehar izan dituen aldaketak, arazoak, ondorio pertsonalak... azalduko dira.

Memoria

Ataza honetan, proiektuaren dokumentua osatuko da.

ATAZA	ESTIMATUTAKO DENBORA (ORDUAK)
Planifikazioa	25 ordu
Aurrekarien prestaketa	15 ordu
Analisia	25 ordu
Diseinua	40 ordu
Inplementazioa	40 ordu
Ebaluazioa	15 ordu
Ondorioak	15 ordu
Memoria	80 ordu
Jarraipen eta Kontrola	25 ordu
Aurkezpena	20 ordu
Guztira	**300 ordu**

Irudia 2.2: Atazen ordu kopurua

Jarraipena eta Kontrola

Ataza honetan proiektuko helburu eta mugarri guztiak betetzen direla bermatuko da. Ustekabekorik gertatuz gero, proiektuak aurrera jarraitzen duela kontrolatuko da. Horretarako zenbait jarraipen bilera burutuko dira. Ataza hau proiektuaren bizitza-ziklo osoan zehar egongo da aktibo.

Aurkezpena

Ataza honetan, proiektuaren defentsarako aurkezpena egingo da. Hau proiektuaren azken atala izango da. Defentsa *Konputazio* espezialitateko epai mahaiaren aurrean egingo da, hau baita GAP honi dagokion espezialitatea.

Ataza hauentzat guztientzat denbora estimazio bat egin da eta 1.2 irudiko taulan ikusi daiteke.

Taulari dagokionez, ordu kopuru handiena memoria osatzen igarotzea estimatzen da. Ataza ugari zehazten direnez proiektuan, atazen artean orduak malgutasunez daude banatuta. Hori bai, diseinuan eta inplementazioan beharko dira ordu kopuru gehien, hauek direlako printzipioz atazik konplexuenak.

2.2.3 Emangarriak

Proiektuan hurrengo emangarriak identifikatzen dira:

Memoria

Proiektuaren inguruko zehaztasun guztiak biltzen dituen dokumentua da. Bertan proiektuaren deskribapena, aurrekarien azterketa, plangintza, garapena, emaitzak eta ondorioak azalduko dira.

Proiektuaren Helburuen Dokumentua

Memoriaren zati bat da, proiektuaren hasieran garatua. Zati horretan, bai proiektuaren garapenaren planifikazioa nolakoa izango den proiektuari buruzko informazioa, baita, honen helburuak ere, azaltzen dira.

Aurkezpena

Proiektua amaitzean, aurkezpen bat egingo da proiektua defendatu ahal izateko. Aurkezpen hau epaimahai baten aurrean egingo da eta proiektuaren nondik norakoak azaldu beharko dira. Azalpen honetan, proiektua azaldu beharko da eta hau zertan datzan.

2.2.4 Mugarriak

Barne-mugarriak

- **2015/03/10,** Egun horretarako Proiektuaren Helburu Dokumentatua egina egon beharko da. Proiektuaren Helburu Dokumentatuan, planifikazioak egina egon beharko du eta horrez gain, atazak definituta izan beharko dute.

- **2015/03/17,** Egun horretarako *Aurrekarien prestaketa* egin beharko da. *Aurrekarien prestaketa*-n *Freeling-med* analizatzaile linguistikoa eta *KAF* fitxategi motak zer eta nolakoak diren azaldu beharko da. Prestaketa hau proiektuko probekin hasi baino lehen egin behar da.

- **2015/03/27,** Martxoaren 27rako *Analisia* egin beharko da. Analisi honetan, *Word Sense Disambiguation* zer den eta honetarako dauden tekniken bilaketa bat egin beharko da. Bilaketaz gain, bi teknika aukeratu beharko dira probak egin ahal izateko beraiekin.

- **2015/04/17,** Apirilaren 14rako *Diseinua* atazak egina egon beharko du.

- **2015/05/05,** Maiatzaren 5rako *Inplementazio* atala osatua egon beharko da. Bertan, aukeratutako teknikak aplikatzeko programak garatuko dira.

- **2015/05/13,** Maiatzaren 13rako *Ebaluazio*-aren ataza egina egon beharko da. Ataza honetan, aukeratu diren teknikekin eman diren proben emaitzak konparatu eta ebaluatuko dira.

- **2015/05/20,** Maiatzaren 20rako *Ondorioak* ataza egingo da, non, proiektuaren ondorio orokorrak azalduko diren.

- **2015/06/12,** Ekainaren 12rako proiektuaren memoria osatuko da.

Kanpo-mugarriak

- **2015/06/17** Egun hau proiektua entregatzeko azken eguna izango da. Hau dela eta, egun honetarako proiektuaren ataza guztiak beteta izan beharko dira, aurkezpena izan ezik.

- **2015/07/8-10** Egun tarte honetan proiektuaren defentsarako aurkezpena egingo da. Egun honetarako, aurkezpenak eginda izan beharko du.

2.2.5 Gantt-diagrama

Hau da, proiektuaren mugarriak kontuan izanik, *Gantt-Diagrama*:

Id.	Ataza	Hasiera	Bukaera	Urtarrila 2015	Otsaila 2015	Martxoa 2015	Apirila 2015	Maiatza 2015	Ekaina 2015	Uztaila 2015
1.	Planifikazioa	2015/01/28	2015/03/10							
2.	Aurrekarien prestaketa	2015/03/10	2015/03/17							
3.	Analisia	2015/03/17	2015/03/27							
4.	Diseinua	2015/03/27	2015/04/17							
5.	Inplementazioa	2015/04/17	2015/05/05							
6.	Ebaluazioa	2015/05/05	2015/05/13							
7.	Ondorioak	2015/05/13	2015/05/27							
8.	Memoria	2015/05/27	2015/06/12							
9.	Jarraipen eta Kontrola	2015/01/28	2015/06/17							
10.	Aurkezpena	2015/06/17	2015/07/08							

Irudia 2.3: Estimatutako Gantt-diagrama

2.3 Lan Metodologia

Bilerak

Bilerak, ikaslearen eta tutorearen artean hitzartuak izango dira. Ez dira finkaturik egongo, hau da, barne-mugarri bakoitza bukatzean bilera bat egongo da. Horrez gain, barne-mugarriren batean arazoren bat izanez gero, bilerak hitzartu ahal izango dira proiektuan aurrera egin ahal izateko. Bilera hauek posta elektronikoz adostuko dira. Arazoak ere posta elektronikoz galdetu eta ahal izanez gero, ebatziko dira.

Planifikatutako ordutegiak

Hasiera batean planifikatutako orduak behar den bezala errespetatzen saiatuko da ikaslea. Baina, aipatu beharra dago, ikaslea lau ikasgaietan matrikulatua dagoela. Hori dela eta, ikasgai hauei lehentasun handiagoa emango zaie. Beraz, honek planifikazioan aldaketak egitea behartu dezake.

Prestakuntza

Tutoreak proposatutako zenbait aurrekari aztertuko dira, eta horietatik ikasitakoa, proiektuan ezartzea saiatuko da.

Garapena

Prestakuntzan ikasitakoa praktikan ezartzea izango da helburua.

2.4 Bideragarritasuna

Proiektua aurrera eramateko beharrezkoak diren baldintzak aztertu ondoren, proiektuaren bideragarritasuna bermatzen saiatuko gara.

1. **Baliabideen kostua.** Proiektuan beharrezkoak diren baliabideak doakoak direla bermatu da.
2. **Baliabideen funtzionamendu bermea.** Erabiliko diren baliabideak proiektuaren garapenean prest eta atzigarri egongo direla bermatu da.

3. **Denbora.** Proiektua aurrera eraman ahal izateko denbora izango dela bermatzen da.

4. **Komunikazioa.** Ikasle eta tutorearen artean komunikazio eraginkor bat izango da.

2.5 Arrisku-analisia

Proiektuaren zehar, hainbat arrisku eta oztopo izango dira, eta hauek proiektuaren arrakasta baldintzatu dezakete. Hori dela eta, arrisku hauek identifikatu behar dira lehenik eta ondoren, hauen aurrean hartu beharreko neurriak hartu beharko dira.

2.5.1 Identifikaturiko arriskuak

Hauek dira proiektuan zehar eman daitezkeen arriskuak:

1. Proiektuan zehar gerta daiteke atazaren bat osatzeko arazoak izatea, eta horrek denbora galera handiak ekar ditzake. Arazo horien artean ere gerta daiteke, beste ikasgaien lan-zamarengatik denborarik ez izatea eta proiektuaren atazak atzeratzea. Horregatik, barne-mugetan proiektua bukatzeko eguna, entrega eguna baino 5 egun lehenago jarri da. Arazoren bat balego, denbora luzatzeko.

2. Proiektuaren parte diren datuen edo zati batzuen galera.

3. Pribatutasuna. Hau da, proiektu honetan laburduren desanbiguaziorako erabiliko diren testuak, kasu errealak dira, eta beraz, nik ez dut eskubiderik testu hauek erabiltzeko. Beraz, soluzio bat bilatu behar da testuak erabiltzeko baina osoak erabili gabe. Hau da, testu bakoitzetik, esaldi kopuru zehatz bat hartu ahalko dut soilik, bestela pribatutasuna ez dudalako mantenduko.

4. Eskuzko etiketatzaileen etiketazioa beharko dut nire emaitzekin konparaketak egin ahal izateko. Arrisku bat izan daiteke, etiketatzaileak beraien etiketazioan atzerapenak izatea eta beraz, nire lana atzeratu ahal izatea. Beraz, adi egon beharko dut etiketatzaileen mugimendu eta atzerapenetara.

5. Eskuzko etiketazioan, gizakiak direnez, erroreak izatea posible da. Hau

da, posible da etiketatzaileari eskatu zaizkion laburdura guztiak ez etiketatzea. Giza-errore gisa hartuko dira arazo hauek.

2.5.2 Kontingentzia-plana

Aurreko azpiatalean aipatutako arriskuen aurrean hartu beharreko neurriak hauek izango dira:

1. Proiektuan zehar, lehen adierazi bezala, atazaren batean arazo edo oztoporen bat gertatuz gero, denboraz justu ez ibiltzeko, planifikazio malgu bat sortu da. Modu honetan, arazoen aurrean, lan-plana alda daiteke. Hori egingo da arazoa txikia bada. Arazoa oso handia izanez gero, proiektua irailean entrega daiteke. Planifikazioak hainbat aldaketa izan ditzake proiektuaren garapenan zehar.

2. Proiektuko daturik ezta zatirik ere galtzeko, astero segurtasun-kopia bat egingo da.

3. Laburduren desanbiguaziorako erabiliko diren testuak ezin ditudanez nik irakurri, testu hauetatik soilik esaldi batzuk hartzeko programa bat sortuko dut. Modu honetan, testu bakoitzetik x esaldi hartuko ditut eta denak testu batean sartu. Honela, testuinguru bat izango dut.

4. Etiketatzaileek atzerapenak neure proiektuan aldaketak ekar ditzake, beraz, etiketatzaileekin komunikazioan egon beharko da, ahalik eta modu onean, etiketazioa luzerako ez joateko. Bi etiketatzaile izango dira, beraz, batek kale eginez gero, bestearen etiketazioarekin soilik egin beharko da lan. Batek soilik eginez gero etiketazioa denbora muga baten artean, etiketatzaile horren etiketazioak soilik erabili beharko dira.

Kapitulua 3

Aurrekariak

Proiektuaren garapenean murgildu aurretik, laburduren desanbiguaziorako erabili izan diren teknikak azalduko dira eta proiektuan erabiliko diren hainbat baliabide deskribatuko dira. Adibidez, *Freeling-med* analizatzaile linguistikoa eta *KAF* XML formatua aztertuko dira.

Hizkuntzaren prozesamendua informatika eta hizkuntzalaritza batzen dituen saila da. Hizkuntzaren bidez, pertsona eta makinen arteko komunikazioa errazteko tresna konputazionalak ikertzeaz arduratzen da.

Hizkuntza prozesatzen duen analizatzaile linguistiko bat *Freeling* da.

3.1 Medikuntzako laburduren desanbiguaziorako teknikak

Hizkuntzaren prozesamenduari dagokionean, teknika ugari daude, baina denak ez daude, medikuntza-arloak behar dituen beharretara zuzenduak.

Horretarako, zein teknika posible dauden begiratu eta aztertu behar da. Bi teknika posible azalduko dira hurrengo atalean:

3.1.1 Grafoetan oinarritutako medikuntza laburduren desanbiguazioa

Grafoetan oinarritutako laburduren desanbiguaziorako teknika azalduko da atal honetan. Berak esaten duen moduan, laburduren desanbiguazioa egi-

teko, grafo bat behar da, eta desanbiguazioa, grafoaren erpinen eta hauen erlazioen araberakoa da.

Grafoetan oinarritutako desanbiguazioan, 2 algoritmo bereizten dira: *PageRank* eta *Personalized PageRank*. Lehenengoak, *Random walk* modeloa erabiltzen du, hau da, ausazko pausuak egiten ditu. Erpin bakoitzari probabilitate berak ezartzen zaizkio ausazko pausuak egitean. Bigarrenak berriz, erpinen garrantziaren egitura bat osatzen du, non erpin batzuk garrantzia handiagoa duten kasu batzuetan. Gainera, erpin batzuei probabilitate altuagoa jartzen zaie.

Personalized PageRank erabiltzeko medikuntza laburduren desanbiguazioan, UMLS erabiltzen da eta hau grafo bat bezala da non kontzeptuak erpinak diren eta hauen arteko erlazioa ertzak. Gainera, bi informazio-iturri behar dira: ezagutza-base bat eta hiztegi bat.

Ezagutza-basea kontzeptu eta hauen erlazioak biltegiratzen dituen fitxategia da. Hiztegiak berriz, dokumentuetan aurkitutako esaldi eta hitzak mapeatzen ditu ezagutza-baseko beraien kontzeptu posiblera.

Nahiko sistema fidagarria dela esan daiteke, erlazioetan oinarritzen delako, hau da, textuingurua kontuan izaten delako, desanbiguazioa egiteko.

Honek eman dezakeen arazoa, testuingurua motza izatea da, eta modu horretan, gai ez izatea desanbiguatu ahal izateko.

Grafoetan oinarritutako sistema bat *UKB* da, eta laburduren desanbiguaziorako erabilgarria da.

3.1.2 Ikasketa automatikoa

Ikasketa automatikoa, disziplina zientifiko bat da non sistemak automatikoki ikastea den helburua. *Ikasketa Automatiko*-ari buruz hitz egiteko, erreferentzia bat egingo zaio, *Disambiguation of Biomedical Abbreviations* artikuluari.

Ikasketa automatikoa aurrera eramateko, hiru letraz osatutako 21 laburduraz osatutako corpus bat sortu zuten egileek. Egileek gainera, corpusa sortzeko erabili zituzten testuetan, laburduraren inguruan beti bere hedapena zihoan eta testu hauek, *abstract*-ak ziren, non kasu horietan, oso normala den laburdura agertzen bada, bere hedapen posiblea ondoan joatea.

Corpusa oso handia zenez eta zailtasunak ematen zituenez, 3 ataletan banatu zuten. Laburdura bakoitzeko 100 azalpen zituena, laburdura bakoitzeko 200 azalpen zituena eta laburdura bakoitzeko 300 azalpen zituena. Azkenekoaren

barruan beste biak zeuden. Ikasketa algoritmo ugari aplikatu zitzaizkien 3 corpusei eta honela, ikasketa automatikoa osatzen joan ziren.

Aplikatu ziren algoritmoen artean, VSM (Vector Space Model) ikasketa algoritmoa, SVM (Support Vector Machines) edo Naive Bayes (NB) ikasketa algoritmoak zeuden. Emaitzarik onenak, SVM ikasketa algoritmoa aplikatuz eman ziren.

3.2 Baliabideak

Proiektuan zehar, hainbat baliabide erabiliko dira, proiektua bideragarria izan ahal izateko. Atal honetan, baliabide horiei buruz hitz egingo da.

3.2.1 UKB

UKB (*Graph Based Word Sense Disambiguation and Similarity*) ezagutza-base bat erabiliz, grafoan oinarritutako hitzen desanbiguazioa (*Word Sense Disambiguation (WSD)*) eta erlazio lexikoak osatzeko programen bilduma bat da. *UKB*-k ausazko pausuak (random walks) ematen ditu, adibidez, *Personalized PageRank* izeneko algoritmoa aplikatzean. Ezagutza baseko grafoko erpinak testuinguruaren arabera sailkatzen ditu. Tresna batzuk barne hartzen ditu grafoak sortu ahal izateko. Hainbat ezagutza-base erabil ditzake eta ohikoena *WordNet* da. *WordNet* ingelesezko datu base lexiko bat da.

UKB hainbat zereginetan erabilia izan da:

- Hitzen desanbiguazioan (WSD) WordNet erabiliz (ingelesez).
- Hitzen desanbiguazioa (WSD) hainbat hizkuntzatan:
  - Euskara
  - Bulgariarra
  - Portugesa
  - Gaztelania
- Hitzen desanbiguazioa (WSD) medikuntza-arloan, *UMLS meta-thesaurus* erabiliz.

- Izendatutako entitateen desanbiguazioa *Wikipedia*-ko grafoa erabiliz.

  UKB *Euskal Herriko Unibertsitate*-ko **IXA** taldeak garatua izan da eta eskuragarri dagoen UKB-ren azkeneko bertsioa, *UKB 2.2* da. Proiektu honetan erabiliko dena. Proiektu honetan erabiliko dena, **2.0** bertsioa izango da.

3.2.2 UMLS - SNOMED CT

UKB-rekin lan egiteko, UMLS (Unified Medical Language System) eta SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) erabiliko dira domeinua medikuntzakoa izango delako, eta horretarako, UMLS da ezagutza-base egokiena. SNOMED CT, lehenengoaren gazteleraz dago eta UMLS (ingelesez)-ren barnean dago, eta *SNOMED CT* erabili nahi dugunez, UMLS ere erabili beharko da.

- **UMLS** UMLS ordenagailu sistemen artean elkareragingarritasuna ahalbidetzen du, eta bertan, osasun eta biomedikuntza hiztegi eta estandarrak bateratzen dituen software eta fitxategi multzoa da.

  Osasun-erregistro elektronikoak, sailkatze-tresnak, hiztegiak eta hizkuntza-itzultzaileak hobetu edo garatzeko erabili daiteke UMLS.

  UMLS-ren erabilera bat osasun informazioa, osasun terminoak, botikaizenak eta fakturazio-kodeak hainbat ordenagailuetan zehar lotzea da. Hainbat adibide daude:

  - Mediku, farmazia eta aseguru-etxetako termino eta kodeak lotzea.
  - Gaixo baten arretarako hainbat departamenturen artean informazioa lotzea.

  UMLS-k gainera, hiru tresna biltzen ditu:

  - **Metathesaurus:** Hainbat hiztegietako kodeak eta terminoak ematen dira.
  - **Sare semantikoa:** Kategoria zabalak (mota semantikoak) eta hauen arteko erlazioak (erlazio semantikoak) ematen dira.
  - **SPECIALIST lexiko eta tresna lexikoak:** Hizkuntza natural prozesatzeko tresnak ematen dira.

  Tresna lexikoak eta sare semantikoa *Metathesaurus* lortzeko erabiltzen dira. Hauez gain, UMLS-ek bere barnean, SNOMED CT du.

- **SNOMED CT**

  Munduan garrantzia, doitasun eta zabalera handien duen terminologia kliniko integral, eleanitz eta kodifikatua da. *SNOMED CT* produktu bat da, zeinaren bitartez datu klinikoak kodifikatu, berreskuratu, komunikatu eta aztertu daitezkeen. Modu horretan, SNOMED CT erabiliz, osasun-profesionalek informazioa modu egoki eta zehatzean azaldu dezakete.

  SNOMED CT kontzeptu, erlazio eta deskribapenen bidez osatua dago. Elementu hauen helburua ezagutza klinikoa eta informazio zehatza erakustea da.

  SNOMED CT *International Health Terminology Standards Development Organisation (**IHTSDO**)* bidez dago banatua. Espainia erakunde honetan dago, eta hau dela eta, Espainiako erakunde eta enpresek SNOMED CT doan erabili dezakete.

3.2.3 Freeling-med

Freeling hizkuntza-analizatzaile bat da. Hainbat hizkuntza analizatzeko gai den kode irekiko liburutegia da. Proiektu honetan, gaztelerarentzat analisi morfologikoa egiteko erabiliko dugu tresna. Hau da, proiektu honetan, testuak gazteleraz egongo dira, eta horiek analizatzeko, Freeling-med erabiliko da.

Freeling-ek hainbat analisi mota egiten ditu: tokenizazioa, kategoria gramatikalen analisia, sintagmen analisia, terminoena eta mendekotasun-zuhaitzena.

Freeling helburu orokorreko analizatzailea medikuntzaren alorrera egokitu zen. *Freeling-med* edozein testu emanik, besteak beste, testu horretako medikuntza-terminoak aurkitu edo detektatzeko gai den tresna da. Horretarako, *Freeling-med*-ek hainbat hiztegi ditu integratuta, adibidez, *SNOMED CT, BotPLUS...*

- **SNOMED CT**

  Dokumentazio kliniko eta informeen adierazkortasuna eta eduki klinikoa eskeintzen duen terminologia kliniko integrala da (ikus 3.2.2 atala). Ondo egituratuta eta zabala den terminologia klinikoa da.

- **BotPLUS**

Espainian merkaturatuta dauden sendagaien izenak gordetzen dituen farmaziaren domeinuko datu-basea da.

- **CIE-10**

  Gaixotasunen Nazioarteko Sailkapenaren 10. bertsioa da. Osasun-txostenen amaieran termino batzuk idazten dira, txostenean agertzen diren gaixotasun nagusiak azalduz eta sailkapen hauen bitartez, termino diagnostiko horiei kode bat/batzuk esleitzen zaizkie.

Horrez gain, *Freeling-med*-en bi prozesatze mota daude: morfologikoa eta sintaktikoa alde batetik, eta bestetik semantikoa. Semantikoa, Freeling-eko analisi-fase guztiak amaituta egiten da aparteko fase batean eta osasun-termino bakoitzari bere klase semantikoa (gaixotasuna, gorputz zatia den...) ematen zaio.

Beraz, orokorrean esan daiteke sistema honen helburua, gazteleraz dauden testu-klinikoetan medikuntza entitateak identifikatzen dituen analizatzaile bat osatzea dela.

3.2.4 Kyoto Annotation Framework (KAF)

Corpus bati, *Freeling-med* aplikatzean, honek, emaitza *KAF* (Bosma et al. 2009) fitxategi batean bueltatzen du. Fitxategi mota honek, bere ezaugarriak ditu.

Kyoto Annotation Framework (KAF) testuak linguistikoki etiketatzeko formatu bat da. Mota honetako fitxategietan testuen analisia gordetzen da, bai maila morfologiko, sintaktiko eta semantikoan ere. Formatuari dagokionean, KAF *XML* (eXtended Markup Language) formatuan oinarrituta dago. KAF fitxategiak honako egitura izaten du: *XML* formatua jarraitzen duenez lehenik *XML* goiburukoa dator eta bertan, *XML* bertsioa, kodeketa mota zehazten da eta horrez gain ere, hizkuntza. Kasu honetan, gaztelerazko testuak analizatuko direnez, hizkuntza gaztelera izango da. Honen jarraiki dokumentuaren erro nodoa irekitzen da, *KAF* bezala izendatua.

```
<?xml version="1.0" encoding="UTF-8"?>
<KAF xml:lang="es">
  <kafHeader>
    <linguisticProcessors layer="terms">
      <lp name="Freeling" version="3.0" timestamp="2015-02-26T18:28:22Z"/>
    </linguisticProcessors>
  </kafHeader>
```

Goiburua eta gero, testuko *WordFormak* datoz, hau da, hitzak testuan agertzen diren moduan. Hauek *wf* etiketarekin zehazten dira. *WordForm* bakoitzak hitz bakoitza zenbatgarren esalditik lortuko den adierazten du *sent* atributuaren bitartez. Horrez gain, *offset* atributuak hitza esaldiko zein karakteretan hasten den argitzen du eta *lenght*-ek berriz zein luzera duen hitz horrek.

```
<wf wid="w172" sent="18" offset="0" lenght="2">FC</wf>
<wf wid="w173" sent="18" offset="2" lenght="1">:</wf>
<wf wid="w174" sent="18" offset="4" lenght="2">80</wf>
<wf wid="w175" sent="18" offset="7" lenght="3">lpm</wf>
```

WordForm-ak eta gero datozen nodo motak, *term* motakoak dira. Hauek termino bat sinbolizatzen dute. Termino bat hitz bat bakarrik edo hitz anitzeko unitate bat izan daiteke, adibidez *edema palpebral* hitz bat baino gehiagoko termino bat izango litzateke. Termino bakoitzean aurkezten den informazioari dagokionez bere lema eta eta kategoria (pos edo part of speech) agertzen dira. Lema hitz baten forma normala da, hau da, hitz baten erroa. Kategoria, aldiz, hitzak esaldiaren testuinguruan duen funtzioa da, hau da, hitza aditza, izena, adjektiboa, izenordaina... izan daiteke. Lema batzuetan ere, erreferentziak agertzen dira. Erreferentzia hauek *SNOMED CT* moduko hiztegietara direnak dira. SNOMED CT biltegi terminologikoan duten identifikadorea eta UMLSn izango luketeen CUI(Concept Unique Identifier)-a ematen da. Baina, hau soilik, *Freeling-med*-ek irteera gisa ematen dituen KAF fitxategietan soilik gertatzen da, KAF normaletan ez da gertatzen.

Lehen esan moduan, *Freeling-med*-ek itzultzen du fitxategi hau eta laburdurak identifikatuz gero, *term*-en bere esanahia esleitzen dio. Hori bai, Freeling-med ez bada gai termino anbiguo bat desanbiguatzeko, esanahi guztiak jarriko ditu KAF fitxategian.

Aurreko irudiko kasua hartuta, *FC* laburdurak, bi esanahi posible ditu:

1. *Fase crónica (tid=t132-0)*
2. *Frecuencia cardíaca (tid=t132-1)*

FC-rentzat bi esanahi ditu, baina *lpm*-rentzat Freeling-ek honako esanahi hau ematen du: *Latidos por minuto*. Hau honela egiten da, lpm ez delako anbiguoa. Hona hemen, aurreko irudiari dagokion terminoanalisiak:

```
<term tid="t132-0" lemma="fase_crónica" pos="NC00YL0">
  <span>
    <target id="w172"/>
  </span>
  <externalReferences>
    <externalRef resource="SCT_es_INT_20130430" reference="278177007" reftype="calificador">
      <externalRef resource="UMLS-2010AB!" reference="C0457343"/>
    </externalRef>
  </externalReferences>
</term>
<term tid="t133" lemma="80" pos="Z">
  <span>
    <target id="w174"/>
  </span>
</term>
<term tid="t134" lemma="latidos_por_minuto" pos="NC00YL0">
  <span>
    <target id="w175"/>
  </span>
  <externalReferences/>
</term>
<term tid="t132-1" lemma="frecuencia_cardiaca" pos="NC00YL0">
  <span>
    <target id="w172"/>
  </span>
  <externalReferences/>
</term>
```

Bestale, baliteke ez aurkitzea laburdurarentzat esanahirik:

```
<wf wid="w138" sent="9" offset="25" lenght="4">EADB</wf>
```

Kasu honetan, *eadb* laburdurarentzat ez du aurkitu esanahi zehatzik:

```
<term tid="t113" lemma="eadb" pos="NCMS000">
  <span>
    <target id="w138"/>
  </span>
  <externalReferences/>
</term>
```

KAF-ek hurrengo nodo gisa *chunk*-ak ditu. *Chunck* bat termino multzo bat da, esaldian funtzio sintaktiko bat izan dezakeena. Chunk analisia nolabaiteko analisi sintaktiko partzial bat da. Chunk bakoitzean aditz edo izen sintagma bat den erakusten zaigu phrase atributuaren bidez.

```
<chunk cid="c58" head="">
  <!--pupilas isocóricas fotorreactivas-->
  <span>
    <target id="t156"/>
    <target id="t157"/>
    <target id="t158"/>
  </span>
</chunk>
```

Azkenik, *deps* nodoa geratzen da. Hemen, laburduren arteko dependentzia erlazioak errepresentatzen dira. Baina, atal hau ez da garrantzitsuena proiektu honetan.

3.2.5 BRAT - Eskuzko etiketazioa

BRAT etiketatzaileek eskuz etiketatu ahal izateko erabiltzen duten sistema da.

BRAT testuen eskuzko anotaziorako erabiltzen den web-oinarria duen tresna bat da. Hau da, existitzen diren testuei oharrak ezartzea ahalbidetzen duen tresna da.

BRAT orokorrean, anotazio egituraturako dago disenatua, non anotazioek ez duten forma libre bat baizik eta forma finko bat duten, automatikoki ordenagailu baten bidez prozesatu eta interpretatua izan daitekeena.

Ondorengo irudia adibide sinple bat da, non sententzia bat anotatua izan den zenbait hitzen entitateak eta beraien moten aipamenak identifikatzeko.

Irudia 3.1: BRAT-1

Etiketatzaileek sistema honen bidez, laburdura bakoitzeko fitxategian, laburdura azaltzen den heinean, hau etiketatuko dute beraien ustetan laburduraren hedapena denarekin. Sistema honen bidez, osatuko dituzte bi etiketatzaileek corpusaren eskuzko etiketazioak, ondoren GoldStandard gisa erabiliko direnak.

BRAT-ek fitxategi bat jasotzen du fitxategi-etiketatu bakoitzeko, non bertan, laburdura agertzen den karaktere kopuruan, honek duen hedapena agertzen den etiketatua.

BRAT-1 irudian ikusten den moduan, BRAT-en laburdurak etiketatuak daude, eta hauei, bere hedapen posibleen artean, bat esleitzen die, *Contexto*-rekin batera dijoana.

Aurrekarien atal honetan, BRAT azaldu da, eskuzko anotazioa egingo duten bi etiketatzaileek sistema hau erabiliko dutelako laburduren eskuzko analisia egin ahal izateko.

Bestalde, proiektu honetan, UKB erabiltzeko asmoa dago eta honek, medikuntza domeinuan erabili ahal izateko, UMLS erabili behar da, grafo bat sortzeko. Hau da, UKB-k 3 elementu behar ditu bere erabilerako: Grafo bat, hiztegia eta kontestua. UMLS-ren bidez, grafoa lortuko da.

Bestalde, kontestua eratu ahal izateko, corpusari, *Freeling-med* aplikatu behar zaio eta hortik, lortzen diren KAF fitxategien bidez, kontestua sortzen da.

Gainera, UMLS-k CUI identifikadoreak ditu termino bakoitzarentzat, eta hauek hiztegia sortzen dute. Hau da, UKB-k erabiliko duen hiztegian, termino bakoitza bere CUI identifikadorearekin joango da.

Kapitulua 4

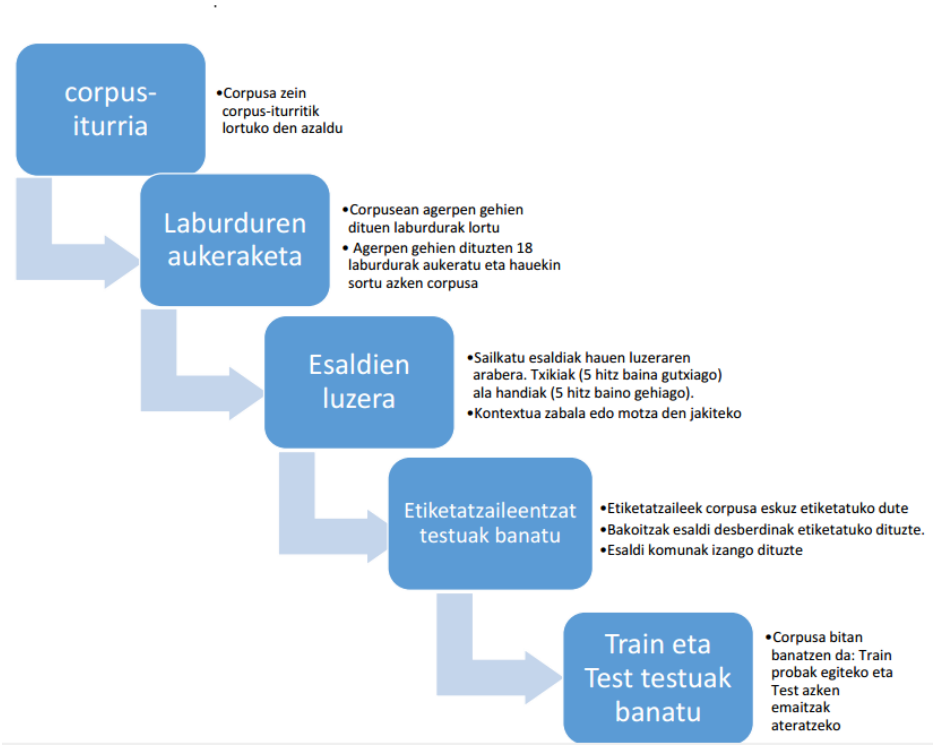
Diseinua

Diseinuaren atala bi zati nagusitan banatuta dago, materialak edo baliabideak alde batetik (4.1 atala) eta metodoak edo teknikak bestetik (4.2 atala).

4.1 Hizkuntza Baliabideak (Materiala)

Atal honetan, corpus-iturria zein izan den, corpusaren eraketa (diseinua) eta Gold Standard-aren eraketari buruz hitz egingo da. Proiektuaren disenuaren arloa azalduko da hemen.

Hizkuntza baliabideen fluxu-diagrama irudian, corpusaren eraketa nondik eta nola egin den azaltzen da, era laburtuan.



Irudia 4.1: Hizkuntza baliabideen fluxu-diagrama

4.1.1 Corpus-a

Lan honetarako erabili den testu-bilduma edo iturri-corpusa Galdakao-Usansolo ospitalean, 2012. urtean idatzitako osasun-txostenetik aterako da. Corpus anitza da, 400 medikuek inguru idatzitako txostenak baitira eta hainbat alorretakoak: kirurgia, ginekologia... Corpus honetan, 24663 testu eta 11.413.076 hitz daude. Corpus honetatik, azpicorpus bat osatuko da, laburdurei bideratuta. Hau da, laburdura anbiguoz osatuta dauden esaldiez osatutako azpicorpus bat sortuko da. Proiektu honetarako interesa, laburdurak dauden esaldiak aukeratzean dago. Hau da, laburdurarik ez dagoen esaldiak ez dira baliogarriak proiekturako behar den corpusean. Beraz, Galdakao-Usansolo ospitaletik lortutako corpusetik, beste azpicorpus bat sortuko da, non hasierako corpus horretatik laburdurez osatuta dauden esaldiak ausaz idatziko diren. Modu horretan, ezingo dira gaixoaren txostenak ikusi eta pribatutasuna mantenduko da.

Azpicorpus honetatik, soilik laburdura batzuk landuko dira, hau da, ez da laburdura guztien desanbiguazioa landuko proiektu honetan. Azpicorpusean agerpen gehien duten laburdurak dira landuko direnak (Svenson et al.)-en

bezala.

Horretarako, azpicorpusetik hainbat datu aterako dira:

- **Agerpena duten laburdurak**

  Testu fitxategi guztietan agertzen diren laburdura guztiak gorde egingo dira eta baita, hauek zenbat agerpen izan dituzten ere. Laburdura eta agerpen kopuruak, beste fitxategi batean gordeko dira, agerpen kopuru gutxien dituztenetatik gehien dituztenetara, honako adibidean bezala:

  HBV 1

  XR 2

  ...

  AC 5

  LDH 6

  ...

  HTA 100

  ...

  Laburduren agerpen kopurua zein den jakitea garrantzia handikoa da proiektu honetan, horiekin egingo baitugu lan. Beraz, zerrenda horretatik aukeratu beharko dira laburdura egokiak lan egiteko.

  Gerta liteke ere, laburdurak detektatzerakoan 3 letra baino gutxiagoko hitzak hartzen dituenez, konjuntzioak edo laburdurak ez diren hitzak hartzea. Kasu horiek, aurrerago ezeztatuko dira.

- **Esaldien batazbesteko luzeera**

  Fitxategietatik laburdurak eta hauen agerpen kopurua ateratzeaz gain, laburdura agertzen den esaldi bakoitzak zenbateko luzera duen ere gordeko da. Datu hau oso interesgarria izango da, gehienbat UKB probatzerakoan, oso ezberdina delako testuinguru handia duen laburdura desanbiguatzea edo testuinguru motza duen laburdura bat desanbiguatzea.

- **Laburdura guztiekin fitxategi berri bat sortu**

  Fitxategien datu orokorrak ere gorde egingo dira, ondorengo pausuetan, corpus-a osatzean, errazago egiteko laburduren agerpena duten esaldien

aukeraketa. Fitxategi honetan, laburdura baten agerpena duten esaldi guztien datuak txertatuko dira hurrengo moduan:

Laburdura + Esaldiaren luzera + Esaldia + Fitxategia

Lehenengo zutabean, agertu den laburdura azaltzen da eta bigarren zutabean, agerpena azaltzen den esaldiaren luzera. Hirugarren zutabean berriz, esaldia bera idatziko da eta laugarren zutabean, laburdura agertu den fitxategiaren izena joango da.

Modu honetan, zein laburdurekin lan egingo dugun jakitean, fitxategi honetan zehar, bilatuko dira bai laburduraren esaldiak eta baita esaldiaren luzeera ere.

4.1.2 Laburduren aukeraketa

Laburdurak eta hauen agerpen kopurua agertzen den fitxategia erabiliko dugu gehien erabili diren 18 laburdurak aukeratu ahal izateko. Aukeraketa egitean, laburdura hautagai moduan onartu dira 5 letra baino txikiagoak diren hitzak. Laburdurak ez direnak ekiditeko, 'stop-words' zerrenda bat definituko da "y, no..."moduko hitzekin, ez daitezen laburdurak kontsideratu. Hori bai, *stop-words.txt* fitxategia erabili arren, 5 letra baino gutxiagoko konjuntzioak eta hitz komunak saihesteko, baliteke, medikuntza-laburdura ez diren hitzen batzuk agertzea (*alta*) eta baita ere, *pH* moduko laburdurak, zeinak soilik hedapen bat duten, *PH*-ren desberdina delako. Beraz, laburdurak bueltatzen dituen fitxategia garbitu egin behar da hitz eta laburdura hauek kendu ahal izateko.

Garbiketa honela egingo dugu:

1.- "Beste"funtzio bat duten hitzak ezabatuko dira:

   (a) *alta* 21209 aldiz agertzen da, baina hitz arrunta da.

   (b) *o* 18823 aldiz agertzen da, baina konjuntzioa da.

   (c) *e* 13987 aldiz agertzen da, baina konjunztioa da, hurrengo adibidean bezala: "hipertensión arterial e hipercolesterolemia".

   (d) *le* 13845 aldiz agertzen da, baina "se le da"gisako esaldietan agertzen da. Beraz, hau ere kendu egingo da.

2.- Anbiguo ez direnak baztertuko dira:

(a) *tto* SEDOM-en arabera, beti da "tratamiento". Beraz, hau ere fitxategitik ezabatu egingo da.

(b) *INR* 6659 aldiz agertzen da, baina, SEDOM-en arabera, beti da *International normalised ratio*. Beraz, hau ere kanpora joango da.

(c) *RDW* 6183 aldiz agertzen da, baina SEDOM-en arabera bere esanahia *Red cell distribution width* da.

(d) *Hb* 3741 aldiz agertzen da, baina SEDOM-en arabera bere esanahia *Hemoglobina* da.

(e) *RS* 3596 aldiz agertzen da, baina SEDOM-en ez da anbigio, *Ritmo sinusal* esanahia bakarrik duelako.

(f) Na 3458 aldiz agertzen da, baina SEDOM-en *Sodio* esanahia bakarrik du, beraz anbiguoa ez da.

(g) *CPK* 3455 aldiz agertzen da, baina bere esanahia bakarra *Creative phoshokinase* da.

(h) *Hgb* 3007 aldiz agertzen da, eta bere esanahia bakarra *hemoglobina* da.

(i) *CrP* 2850 aldiz agertzen da eta bere esanahia bakarra *creatinina plasmática* da.

(j) *Rx* 2833 aldiz agertzen da, baina bere esanahia bakarra *radiografia* da.

(k) *EPOC* 2630 aldiz agertzen da, baina bere esanahia bakarra *Enfermedad pulmonal obstructiva crónica* da. Beraz, hau ere ezabatu egingo da.

(l) *RMN* 2345 aldiz agertzen da, baina bere esanahia bakarra *Resnonancia magnética nuclear* da.

(m) *TnT* 2246 aldiz agertzen da eta bere esanahia bakarra *Troponina T* da. Beraz, hau ere ezabatuko da.

(n) *Hto*-k 4511 agerpen ditu horrela idatzita, baina bere esanahia bakarra da: *Hematócrito*. Beraz, ez da anbiguoa.

3.- Minuskulak eta maiuskulen arabera, anbiguo ez direnak ezabatu egingo dira:

(a) PH-k bi hedapen posible ditu: *Paciente hipertenso* eta *Progenitores hematopoyéticos*. Baina, fitxategian dagoen laburdura ez

da PH baizik eta **pH**. Kasu honetan hedapenak ez dira berdinak eta hedapen kopurua, bakarra da. Hedapena *acidez o alcalinidad de una solución* da eta beraz ez da anbiguo. Hau dela eta, **pH** fitxategi honetatik ezabatuko da.

(b) CM anbiguoa da eta 4 hedapen posible ditu, baina fitxategian **cm** azaltzen da 3810 agerpenekin, eta kasu hau ez da anbiguoa, bere hedapen posible bakarra *centímetro* delako.

(c) MM-ren kasua CM-ren berdina da, MM-k 4 hedapen posible dituelako. Baina, fitxategian, **mm** da azaltzen dena 16505 agerpenekin, baina hauen hedapen posible bakarra *milímetro* da eta beraz, ez da laburdura anbiguo bat.

(d) HG-k bi hedapen posible ditu: *Hemorragia gástrica* eta *Hipogastrio*. Baina, fitxategian azaltzen den laburdura, **Hg** da 2878 agerpenekin, baina honek duen hedapena bakarra da : *mercurio*. Beraz, laburdura hau ez da anbiguoa eta fitxategitik ezabatuko da.

Garbiketa hau egin eta gero, fitxategiaren bukaeran, agerpen kopuru altuenak dituzten laburdura anbiguoak egongo dira eta horietatik 18 aukeratuko dira. 18 laburdura hauek, 2000 agerpen-kopuru baina gehiago dituzten laburdurak izango dira. Hau da, 2000 agerpenen baina gehiago dituzten, eta garbiketaren ondoren geratu diren laburdurak dira. Aukeraketa eginda, hauek dira, agerpen kopuru altuena duten eta beraz, probetarako erabiliko diren 18 laburdurak:

Taula 4.1: Laburdurak

LABURDURAK	**Agerpen kopurua**
CD	2340
EAC	2376
DA	3196
TP	3414
VI	3460
K	3561
ECG	3617
FA	3700
QRS	4032
h	4269
PCR	4614
DM	4921
T	5037
FC	5869
C	6351
TA	7839
TAC	8060
HTA	8574

4.1.3 Sailkapena esaldien luzeraren arabera

Corpus-a eratzeko erabiliko diren esaldi guztiak ez dira tamaina berdinekoak. Gure susmoa, testuingurua txikia dutenean teknika batzuk okerrago ibiliko direla da, testuinguru handia dutenean baino. Hau dela eta, luzera ezberdinko esaldiekin proba egin nahi izan dugu. Argi izan behar den beste ezaugarri bat, esaldiak "salto de línea-ka daude banatuta, beraz, esaldiak hainbat esaldi izan ditzake eta esaldi bakoitzean, hainbat laburdura-agerpen egon daitezke.

Hori dela eta, laburdura bakoitzeko direktorio bat sortuko da, non 3 fitxategi ezberdin gordeko diren. 3 fitxategi horien izenak *denak.txt*, *txiki.txt* eta *handi.txt* fitxategiak izango dira.

1. *denak.txt* : Fitxategi honetan, laburdura bakoitzak dituen esaldi guztiak gordeko dira. Adibidez, **HTA** laburduraren esaldi guztiak, HTA direktorioan dagoen *denak.txt* (*/HTA/denak.txt*) fitxategian gordeko da.

2. *txiki.txt* : Fitxategi honetan, 5 hitz baina gutxiagoko luzera duten eta laburdura barnean duten esaldiak gordeko dira. Adibidez : **No HTA** edo **No DM**. Hauek laburdura bakoitzaren izena duen direktorioaren barnean dagoen *txiki.txt* fitxategian gordeko da.

3. *handi.txt* : Fitxategi honetan berriz, alderantzizkoa egingo da. Hau da, 5 letra baina gehiago dituzten eta laburdura bertan duten esaldiak gordeko dira fitxategian.

Ondorengo taulan, *Laburdurak-1* izeneko taulan, ikusgai dugu laburdura bakoitzeko zenbat esaldi dauden bai "txikiak"eta baita "handiak"ere. Horrez gain, denera laburdurako zenbat esaldi dauden ere ikusgai dago. Honela ikusi ahalko da zein laburdura den esaldi kopuru gehien duena. Probetarako corpusa sortzean, handi eta txikien arteko proportzioa mantenduko da.

Taula 4.2: Laburdurak-1

	TXIKI	**HANDI**	**DENAK**
C	190	2942	3152
CD	13	959	972
DA	12	1857	1269
DM	407	2037	2444
EAC	58	1104	1162
ECG	90	1644	1734
FA	56	1727	1783
FC	26	2831	2857
h	157	1823	1980
HTA	1526	2721	4247
K	4	1765	1766
PCR	67	2206	2273
QRS	4	1887	1891
T	28	2194	2222
TA	76	3697	3773
TAC	401	3366	3767
TP	24	1640	1664
VI	37	1639	1676

Laburdurak-1 taulan ikusten den moduan, esaldi kopuru gehien duen laburdura **HTA** da 4247 esaldirekin eta esaldi gutxien duena berriz, **CD** da 972 esaldirekin. Horrez gain, laburdura bakoitzeko 5 hitz baino gehiago dituen esaldiak gehiago daude bostera iristen ez direnak baino. Bestalde, probak

egiteko, laburdura bakoitzeko esaldi kopuru bera erabiliko da. Esaldi kopuru gutxiena CD laburdurak duenez 972-rekin, 900 izan da laburduren corpusa definitzeko aukeratu dugun esaldi kopurua. Beraz, laburdura bakoitzetik 900 esaldi hartuko dira.

Esaldi banatze honek arazo txiki bat eman dezake. Adibidez, esaldiak 5 hitz baino txikiagoak direnean baliteke, esaldiak adibidez soilik, 2 edo hitz bakarrekoak izatea. Adibide gisa:

No DM

No HTA

Bi adibide hauetan hedapena zein den asmatzea zaila da, laburdurak DM edo HTA kasuan, ia ez dutelako testuingururik. Beraz, txikiak ez ditugu soilik 5 hitz baina gutxiagokoak hartuko baizik eta 3 hitz eta 5 hitz arteko luzera duten esaldiak. Hau da, 3 hitz baino gutxiago duten esaldiak kanpoan geratuko dira.

No DM EZ

No DM tipo 1 BAI

Beraz, programa baten bidez, 3 hitz baino gutxiago dituzten esaldiak kendu egingo dira txikien multzotik. Baldintza hau kontuan hartuz, laburduren kopuruen taula eguneratu da eta *Laburdurak-2* taulan adierazi da.

Taula horretan, ikusten den bezala, laburdura batzuk txikietan esaldi kopuru berarekin geratu dira, adibidez, *CD*-ren kasuan. Baina, beste batzuetan, txikien esaldi kopurua jaitsi egin da eta horrek, beraien esaldi kopuru globala ere jaitsi egin du. Jaitsi diren horien artean, bi laburdura nabarmentzen dira: **HTA** eta **DM**. Lehenengoak ematen du beherapen handiena, 1526 esaldietatik 697 esaldietara jaisten. Horrek esan nahi duenez, eskuragarri genuen corpusean, *No HTA* gisako esaldi ugari zeuden, eta honelako esaldi guztiak kendu egin dira. Gure susmoa, gaixoak "hipertensión arterial"eta "diabetes mellitus"izan ote dituen oso maiz adierazten dela txostenetan, 'No HTA' edo 'No DM' moduan.

Bestalde, **DM** 407 esaldietatik 331-rekin geratu da. Hala ere, *Laburdurak-1* taulan moduan, **CD** da guztira esaldi kopuru gutxien dituen laburdura, 972 esaldirekin, eta esaldi kopuru gehiena duen laburdura kasu honetan, **TA** da 3765 esaldirekin guztira.

Taula 4.3: Laburdurak-2

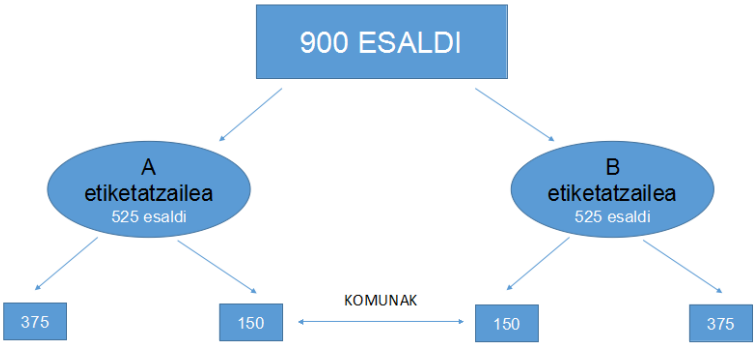
	TXIKI	**HANDI**	**DENAK**
C	188	2942	3130
CD	13	959	972
DA	12	1857	1869
DM	331	2037	2368
EAC	50	1104	1166
ECG	82	1644	1726
FA	38	1727	1765
FC	26	2831	2857
h	157	1823	1980
HTA	697	2721	3418
K	3	1765	1768
PCR	50	2206	2257
QRS	4	1887	1891
T	20	2194	2214
TA	68	3697	3765
TAC	202	3366	3568
TP	15	1640	1655
VI	16	1639	1675

4.1.4 Etiketatzaileentzat testuak banatu

Corpusa osatzeko erabiliko diren laburdurak aukeratuta, esan den moduan, bakoitzetik, 900 esaldi aukeratuko dira. Esaldi guztiak izanda, bi etiketatzaileri banatuko zaizkie, hauek aukeratu ahal izateko, esaldi bakoitzean dagoen laburduraren hedapen egokia eta gure sistemek lortutako datuen aurka konparatuko dira, honela, sistemen asmatze-tasa ezagutzeko.

Modu honetan Gold Standard-a sortuko da. Etiketatzaile hauek farmazia alorrekoak dira. Bata, demagun A deitzen dela eta bestea, B dela. Laburdura bakoitzeko dauden 900 esaldietatik, bi etiketatzaile daudenez, bakoitzari 525 esaldi banatuko zaizkio. 525 esaldietatik, 150 esaldi komunak izango dituzte, hau da, esaldi berdinak, eta beste 375-ak ezberdinak izango dira.Berdinak diren esaldiak konparatuz A eta B etiketatzaileen arteko adostasuna neurtzen da, ingelesez, Inter Annotator Agreement dena.

Ondorengo 3.1 irudian azaltzen den gisa.

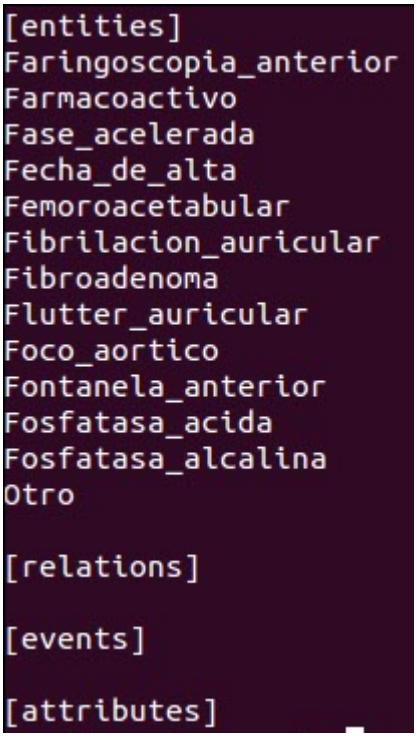


Irudia 4.2: Etiketatzaileen banaketa

Hala ere, arriskuen analisian ikusi den gisa, arriskua izan daiteke etiketatzaileek beraien lana ez bukatzea epe baten barruan edo zuzenean etiketazioa ez egitea.

Esan beharra dago ere, bai A eta B etiketatzaileei esaldiak proportzioan banatuko zaizkiela. Hau da, laburdura batek X txiki eta Y handi baditu, bi etiketatzaileei proportzioan emango zaizkie bai txikiak eta handiak. Proportzio hau, laburduren agerpen errealen araberakoa izango da (ikus 4.2 taula). Ez ditu etiketatzaile batek txiki guztiak izango, baizik eta, 2 etiketatzaileek bai esaldi "txikiak"eta "handiak"izango dituzte.

Etiketatzaileei, corpusa banatzean, laburdura bakoitzeko, zein hedapen posible etiketatu ahal dituzten adieraziko zaie *Annotation.conf* fitxategian. Hedapen posibleak, laburdura bakoitzeko, *Annotation.conf* fitxategian dituzte etiketatzaileek, eta fitxategi horretan dauden hedapen posibleetatik aukeratzen dute hedapen posiblea. Adibidez, FA-ren kasuan, bere *Annotation.conf* fitxategiak *FA annotation.conf* irudiko formatua du:



Irudia 4.3: FA annotation.conf

4.1.5 Train eta Test testuak aukeratu

Corpusan bi atal zehatz daude: **Train** eta **Test**.

- **Train** Atal hau, probak egiteko erabiliko den atala da. Ikasketa automatikoa egiteko erabiltzen da, hau da, probak eta hauen hobekuntzak, corpusaren atal honekin egingo dira. Atal honekin entrenamendu bat egingo dela esan daiteke. Train corpus guztiaren atalik handiena izango da. Kasu honetan, ez da ikasketa automatikoa erabiliko, baina maiztasunak kalkulatzeko erabiliko da.

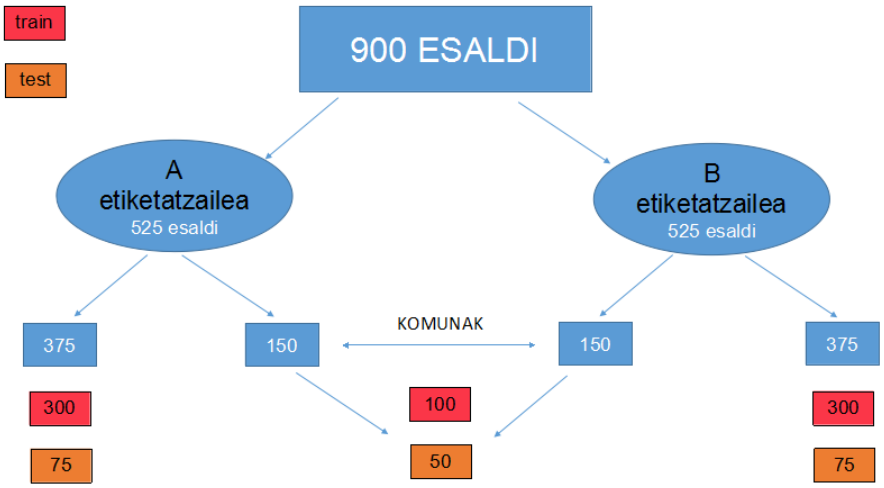
  900 esaldi badaude laburdura bakoitzeko, horretatik 700 esaldi erabiliko dira Train moduan. Horrek esan nahi du, etiketatzaile bakoitzak, train eta test atalak etiketatuko dituela. Hau da, A etiketatzailearen etiketaziotik atal bat Train izango da eta B etiketatzailearen etiketazioaren beste zati bat ere Train izango da. Gainera, komunean dauden esaldietan ere, zatiketa egingo da bai Train eta Test-erako

- **Test**

  Corpus-aren atal ezkutua dela esan daiteke. Hau da, **Test** diren esaldiak, soilik proiektuaren amaieran erabiliko dira azken emaitzak atera ahal izateko. Train entrenamendurako da eta Test berriz, entrenamenduak egin eta hobetu ondoren, azken emaitzak emateko corpus-aren atala da.

Beraz, laburdura bakoitzeko 900 fitxategi-lerro hauek banatu egin behar dira. Kontuan izan behar da, *Train* atalak handiago izan behar duela eta etiketatzaile bakoitzari bi atalak banatu behar zaizkiela. *Train-Test banaketa* irudian ikus daiteke, etiketatzaile bakoitzari zenbat esaldi eta horietatik, zein izango diren *Test* eta *Train*-erako.

Ikusten den moduan, bai A eta B etiketatzaileen 525 esaldietatik 400 Trainerako eta 125 Test-erako erabiliko dira (Bakoitzak dituenetatik 300 Trainera eta komunetatik 100 Test-era).



Irudia 4.4: Train-Test banaketa

4.1.6 Maiztasunak

Proiektu honetan GoldStandard-a sortuko duten 2 etiketatzaileak, beste proiektu bat ari dira etiketatzen. Etiketazio horietaz baliatuko gara, etiketatuta dauden laburduren hedapenak ikusteko. Bi etiketatzaileek etiketatutako corpusetik laburduren agerpenak erauzi dira testu fitxategi batera. Itxura hau du laburduren eskuzko anotazioak dituen fitxategiak:

T7 Grp_Medicamento 623 626 AAS

#1 AnnotatorNotes T7 siglas: ácido acetilsalicico

T8 hipertensión_arterial 780 783 HTA

Fitxategi honetan, azken zutabean laburdura agertzen da eta bere hedapena, bi modutan ager daiteke. Botika bada, 'nota' moduan (ikus T7 adibidea) eta '#1' oharra, eta bestela lerro berean (ikus T8 adibidea).

Etiketazio-fitxategi horrretatik, laburdura bakoitzeko dauden hedapen posibleak eta horrez gain, bakoitza zenbat aldiz izanden aukeratua gordeko da. Honela, laburdura bakoitzeko, hedapen bakoitzak duen agerpenen ehunekoa jakin ahal izango da. Beraz, sistemak erabiltzerakoan, maiztasun hauetaz baliatuko gara.

Etiketatutako laburdura guztiak *maiztasunak.txt* izeneko fitxategi batean gordeko dira. Fitxategi honetan, laburdura bakoitza eta honen hedapen

guztiak bere hautatze-ehunekoarekin agertuko dira, eta horrez gain, *batura* izeneko balio bat egongo da, zeinak kalkulatuko duen laburdura bakoitza corpusean zenbat aldiz agertu den.

Laburdura horien agerpen-ehunekoak etiketatzaileen fitxategian, *maiztasunak.txt* fitxategian aurkituko dira, eta hona hemen *maiztasunak.txt*-ren formatuaren eredu bat:

maiztasunak.txt :

acp:

batura=13

angioplastia coronaria percutánea=0%

arteria cerebral posterior=0%

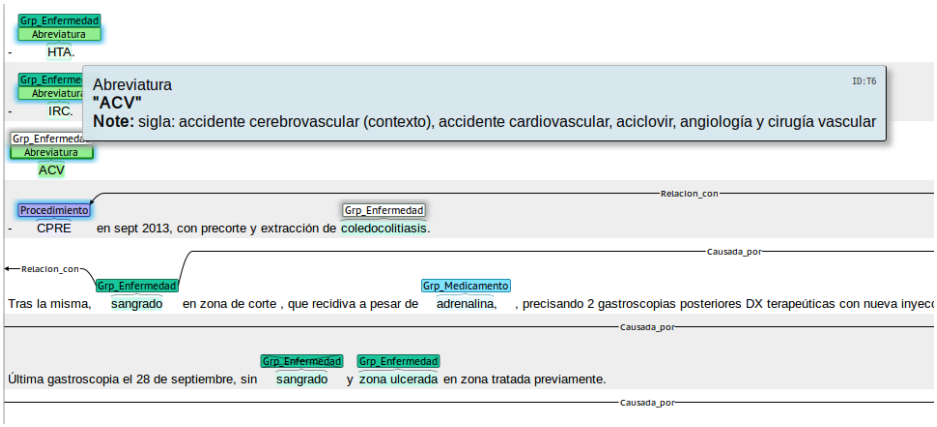
arteria comunicante posterior=0%

auscultación cardiopulmonar=100%

...

Adibidean ikusten den gisa, *maiztasunak.txt* fitxategian, existitzen dira laburdurak, anbiguoak izan arren, erabili den corpusean, hedapen posible bakarra dutenak. Horren adibide argia da **acp** laburdura, zeinak, 4 hedapen posible dituen, baina corpus horretan azaltzen den 13 esaldietan, soilik *auscultación cardiopulmonar* esanahia duen. Honelako hainbat kasu aurkitu dira erabili den corpusean. Baina, kasu hauek aurrerago azalduko dira. *maiztasunak.txt* fitxategi hau, maiztasunen sisteman, Train ataleko fitxategietarako erabiliko da, eta Test-erako, Train-etik aterako dira maiztasunak.

Maiztasunak ateratzeko, *ACV-ren etiketazioa* irudian agertzen den moduan, hedapen posibleetatik, kontaketa, *contexto* duenari egiten zaio. Kasu horretan, ACV laburdura dago ikusgai eta kasu horretan, *accidente cerebrovascular* da hedapen hautatua.



Irudia 4.5: ACV-ren etiketazioa

4.1.7 Laburduren deskribapena

Proiektuan zehar, ikusi den bezala, laburdurak hainbat inguruetan landu dira: Eskuzko etiketatzean ADR corpusean, Freeling-med-en(ikus 3.2.3 atala) hiztegitik, *SEDOM* hiztegitik (etiketatzaileek etiketatzeko dituzten hedapen posibleak), corpus berezituan hedapen zehatz batzuekin... Hauek guztiak izanik, denek berdinak erabiltzeko moldatu behar ditugu, hau da, bakoitzean zein erabiltzen diren ikusirik, hedapen posible batzuk soilik definitu behar ditugu, ebaluazioa egiterakoan sistema guztiek hedapen berak izan ditzaten.

Taulen zutabeak honela daude antolatuta:

- 1.Zutabean *SEDOM* hiztegian laburdurak dituen hedapen posibleak agertuko dira.
- 2.Zutabean *Freeling-med*-ek erabiltzen dituen laburduraren hedapenak agertuko dira.
- 3.Zutabean *Eskuz etiketatutako ADR corpuse*-an dauden laburduraren hedapen posibleak agertuko dira.
- 4.Zutabean Etiketatzaileek corpus honetan laburdura bakoitzeko etiketatu dituzten hedapenak agertuko dira.
- 5.Zutabean aukeratuak izango diren laburduraren hedapenak azalduko dira.

Tauletan ikusiko den moduan, *VI*,*ECG* eta *DM* laburduretan, aukeraketa egitean, soilik hedapen bat aukeratzen da (VI-ren kasuan "otro ere") eta kasu horietan, laburdurak corpusean anbiguo ez direla esaten da.

Etiketatzaileek "otro"etiketatzean, hedapen posibleetan ez dagoen esanahi bat duela esan nahi du.

LAB.	SEDOM	FREELING-MED	ADR corpus	ETIKETATUAK	AUKERATUAK
C	ácido ascorbico caloría grande Canino Carbono cirugía Complemento Consulta Culombio c caloría pequena c Centi Otro	ácido ascorbico, vitamina canino carbono caloría grande cirugía complemento consulta culombio	consulta carbono culombio canino complemento caloría grande cirugía ácido ascórbico, vitamina	estandarizatuak hartu	cabezacuello consciente C Celsius cuerpos cayados
CD	Cluster of differentiation Coito dirigido Colon descendente Coronaria derecha cd Cuenta dedos Otro	Cluster of differentiation Coito dirigido Colon descendente Coronaria derecha	Cluster of differentiation Coito dirigido Colon descendente Coronaria derecha Racimos de diferenciación	Cluster of differentiation Coronaria derecha Otro	Cluster of differentiation Coronaria derecha colon descendente Otro
DA	Dermatitis atópica Descendente anterior Doble anexectomía Ductus arterioso Otro	Dermatitis atópica Descendente anterior Doble anexectomía Ductus arterioso	Dermatitis atópica Descendente anterior Doble anexectomía Ductus arterioso	Descendente anterior Doble anexectomía Otro	Descendente anterior Doble anexectomía Otro

Taula 4.4: Laburduren deskribapena

LAB.	SEDOM	FREELING-MED	ADR corpus	ETIKETATUAK	AUKERATUAK
DM	Dermatomiositis Diabetes mellitus Duramadre Otro	Diabetes mellitus	Dermatomiositis Diabetes mellitus Duramadre dermatomiositis duramadre	Diabetes mellitus	Diabetes mellitus
EAC	endarteriectomía carotidea Enfermedad arterial coronaria Otro	endarteriectomía carotidea Enfermedad arterial coronaria	endarteriectomía carotidea Enfermedad arterial coronaria	endarteriectomía carotidea Enfermedad arterial coronaria	endarteriectomía carotidea Enfermedad arterial coronaria
ECG	Electrocardiograma Enfermedad Escala de Glasgow del coma Otro	Electrocardiograma	Electrocardiograma Enfermedad Escala de Glasgow del coma Electroencefalograma	Electrocardiograma	Electrocardiograma
FC	Frecuencia cardiaca Fase crónica Otro	Frecuencia cardiaca Fase crónica	Frecuencia cardiaca Fase crónica	Frecuencia cardiaca Fase crónica Otro	Frecuencia cardiaca Fase crónica
h	Altura Hematie Hora Otro	Altura Hematie Hora Hospital	Altura Hematie Hora Hemoglobina	Hora Hospital	Hematie Hora Hospital

Taula 4.5: Laburduren deskribapena

Taula 4.6: Laburgarren deskribapena

LAB.	SEDOM	FREELING-MED	ADR corpus	ETIKETATUAK	AUKERATUAK
FA	Faringoscopia anterior Farmacoactivo Fase acelerada Fecha de alta Femoroacetabular fibrilación auricular Fibroadenoma Flutter auricular Foco aortico Fontanela anterior Fosfatasa acida Fosfatasa alcalina Otro	fibrilación auricular Farmacoactivo Fosfatasa alcalina	Faringoscopia anterior Farmacoactivo Fase acelerada Fecha de alta Femoroacetabular fibrilación auricular Fibroadenoma Flutter auricular Foco aortico Fontanela anterior Fosfatasa acida Fosfatasa alcalina Otro	fibrilación auricular Farmacoactivo Fosfatasa alcalina	fibrilación auricular Farmacoactivo Fosfatasa alcalina
HTA	hipertensión arterial diastolica hipertensión arterial Histerectomia total abdominal Otro	hipertensión arterial Histerectomia total abdominal	hipertensión arterial Histerectomia total abdominal	hipertensión arterial Otro	hipertensión arterial Histerectomia total abdominal otro

LAB.	SEDOM	FREELING-MED	ADR corpus	ETIKETATUAK	AUKERATUAK
K	Karnofsky escala de Kelvin potasio Vitamina K k Kilo Otro	Karnofsky escala de Kelvin potasio Vitamina K	Karnofsky escala de Kelvin potasio Vitamina K Termodinámica	Potasio Otro	Kelvin potasio Vitamina K otro
PCR	Parada cardiorrespiratoria Plantar cutaneos reflex Polimerase chain reaction Proteina C reactiva Otro	Parada cardiorrespiratoria Polimerase chain reaction Proteina C reactiva	Parada _cardiorrespiratoria Plantar cutaneos reflex Polimerase chain reaction Proteina C reactiva	Parada cardiorrespiratoria Polimerase chain reaction Proteina C reactiva	Parada cardiorrespiratoria Polimerase chain reaction Proteina C reactiva
QRS	Parte del trazado del electrocardiograma que representa la despolarización ventricular Quiste renal simple Otro	Parte del trazado del electrocardiograma que representa la despolarización ventricular Quiste renal simple	Parte del trazado del electrocardiograma que representa la despolarización ventricular Quiste renal simple	Parte del trazado del electrocardiograma que representa la despolarización ventricular Quiste renal simple Otro	Parte del trazado del electrocardiograma que representa la despolarización ventricular Quiste renal simple

Taula 4.7: Laburduren deskribapena

LAB.	SEDOM	FREELING-MED	ADR corpus	ETIKETATUAK	AUKERATUAK
T	Temperatura Timo torácico Tumor t Tiempo t Translocation Otro	Temperatura Timo torácico Tumor Tiempo t Translocation Total Toxoplasma Transferasa Talla Catéter_tipo T	Temperatura Timo torácico Tumor	Temperatura Timo Tiempo Otro	Temperatura Timo torácico Tumor Tiempo t Translocation Total Transferasa Talla Catéter tipo T
TA	Temperatura ambiente tensión arterial Terminologia anat'omica Tratamiento actual Traumatismo abdominal Otro	Temperatura ambiente tensión arterial Tratamiento actual Traumatismo abdominal	Temperatura ambiente tensión arterial Terminologia anat'omica Tratamiento actual Traumatismo abdominal Histeroctomia total abdominal	tensión arterial Otro	tensión arterial temperatura ambiente otro

Taula 4.8: Laburduren deskribapena

LAB.	SEDOM	FREELING-MED	ADR corpus	ETIKETATUAK	AUKERATUAK
TAC	Tomografia axial computarizada Tratamiento asertivo comunitario Otro	Tomografia axial computarizada	Tomografia axial computarizada Tratamiento asertivo comunitario	Tomografia axial computarizada	Tomografia axial computarizada tratamiento asertivo comunitario
TP	túbulo proximal Tiempo de protrombina Trabajo de parto Transplante pancreatico Transplante pulmonar Trastorno de personalidad Trombopenia Tuberculosis pulmonar Otro	túbulo proximal Tiempo de protrombina Transplante pancreatico Transplante pulmonar Trombopenia Tuberculosis pulmonar	túbulo proximal Tiempo de protrombina Transplante pancreatico Transplante pulmonar Trastorno de personalidad Trombopenia Tuberculosis pulmonar	Tiempo de protrombina Trastorno de personalidad Otro	Tiempo de protrombina trastorno de personalidad trombopenia otro

Taula 4.9: Laburduren deskribapena

LAB.	SEDOM	FREELING-MED	ADR corpus	ETIKETATUAK	AUKERATUAK
VI	ventrículo izquierdo Via intravenosa Volumen de inspiración Otro	ventrículo izquierdo Volumen de inspiración	ventrículo izquierdo Via intravenosa Volumen de inspiración	ventrículo izquierdo Otro	ventrículo izquierdo otro

Taula 4.10: Laburduren deskribapena

4.1.8 Etiketatzaileen arteko adostasuna

Etiketatzaileei bidali zaizkien testuak banatuta bidali dira, laburdura bakoitzeko 525 (train eta test barruan) esaldi eta horietatik 150 esaldi bi etiketatzaileek komunak izango dituzte, hau da, 150 esaldi horiek berdinak izango dira. Hauek, etiketatzaileen arteko adostasuna edo *Interannotator Agreement*-a osatzeko erabiliko dira. Hasiera bateko ideia, *Train* eta *Test* banatzea zen komunak ziren 150 esaldietan. Azkenik, pentsatu da, bi etiketatzaileen adostasuna zein den jakiteko, hobea dela, 150 esaldi komunetan dauden laburdura guztiak konparatzea.

Interannotator Agreement-a erabiltzen da, anotazioaren fidagarritasuna neurtzeko. Era berean, atazaren zailtasuna neurtzeko balio dezake. Hau da, etiketatzaile bakoitzak etiketatu duena, makinak atera duenarekin konparatu aurretik, lehenengo, bi etiketatzaileek etiketatutakoa konparatzen da, bien adostasuna antzekoa dela ziurtatzeko. *Interannotator Agreement* baxua balitz, ez lirake etiketazioak baliokoak izango, ez daudelako ados berdinak diren esaldietan laburduren desanbiguazio egokia egitean.

LABURDURA	**AKORDIO EHUNEKOA**
C	97.802%
CD	98%
DA	98.9583333333333%
DM	100%
EAC	97%
ECG	97.979797979798%
FA	92.7083333333333%
FC	98.989898989899%
h	100%
HTA	95.2941176470588%
K	100%
PCR	97.7272727272727%
QRS	98.989898989899%
T	68.6868686868%
TA	96.8085106382979%
TAC	96.969696969697%
TP	94.1176470588235%
VI	91.25%

Taula 4.11: Inter Annotator Agreement

Inter Annotator Agreement taulako datuak ikusirik (ikus 4.11 taula), argi ikusten da, kasu guztietan, bi etiketatzaileen arteko akordio maila oso altua dela, batzuetan 100%ekoa. Beste kasuetan, etiketatzaile batek, ez dituenez dauden laburdura guztiak etiketatu, ehunekoa jaitsi egiten da, hori akats bezala kontatzen delako. Horiek, giza- akatsak dira, eta normalak dira.

Hala ere, esan beharra dago, bai *C* eta *T* laburduren kasuan, bi etiketatzaileek laburdura bakoitza etiketatzean, *Otro* etiketa esleitu dietela, eta *Otro* horren barnean, beraiei hedapen posiblea zena iruditzen zitzaiena idatzi dute. Bi kasu horietan, *Otro* horren barruan dauden hedapenak konparatu dira eta hortik atera dira ehunekoak. Otro etiketa erabili dute etiketatzaileen ustezko hedapena ez zegoenez anotazioa konfiguratzeko proposaturiko hedapenen artean.

Adibidez, C laburduraren kasuan, biek laburdura agerpen berberari "otro"deizen die, baina batek "Celsius"eta besteak "Centigrado"deritzo hedapen posibleari.

Modu honetan, etiketatzaileen etiketazioak fidagarriak direla kontuan hartu daiteke, beraien arteko adostasuna nahiko altua delako.

4.2 Teknikak (Metodoak)

Proiektu honetan, 3 teknika erabiliko dira: Auzkotasunean oinarritutako teknikak, maiztasunetan oinarritutako teknikak eta UKB, grafoetan oinarritutakoa. Atal honetan, 3 sistema hauek azalduko dira.

4.2.1 Ausazkotasuna - Baseline

Tekniketan beti egoten da bat "baseline"deitzen dena. Hau da, ia esfortzurik gabe, aukeraketa ausaz eginda, zein emaitza lortzen diren jakiteko erabiltzen da teknika hau. Orokorrean, "baselinea-k esfortzu handirik ez duenez behar, emaitza eskasenak emango dituena izango dela pentsatzen da eta erabiliko diren beste teknikek, baseline-n emaitzak hobetuko dituztela pentsatzen da.

Sistema honek, laburdura batek dituen hedapen posible guztietatik, laburdura agerpen bakoitzari ausaz etiketatzen dio hedapen posiblea. Hau da, corpusean laburdura agertzean, honen hedapen posibleetatik, ausaz bat aukeratu eta hori etiketatuko dio bere hedapen gisa. Beraz, sistema honetan, fitxategi bat osatuko da, non testu-fitxategiko laburduraren agerpen bakoitzeko, ondorengo formatua idatziko den:

T1 hipertensión_arterial 373 376 HTA

T2 Histerectomía_total_abdominal 455 458 HTA

T3 hipertensión_arterial 503 506 HTA

Formatu hau BRAT eskuzko etiketatzaileak emaitza moduan ematen duen berdina da. Erabilitako metodo guztiek emaitza formatu honetan emango dute, eskuzkoarekin konparaketa modu automatikoan egin ahal izateko. Teknika honetan laburdura bakoitzerako bere corpusean agertzen diren laburdura guztietarako hedapen bat aukeratzen da 4.4-4.10etan aukeratutakoen artean. Hau da, VI laburdurari dagokion corpusean, VI guztiei 4.10 taulako aukeratuen artean hedapen bat emango zaio.

Aurreko atalean adierazi den bezala, sistema honek 2 fitxategi atearko ditu: A etiketatzailearen datuekin konparatuko direnak eta B etiketatzailearekin konparatuko direnak. Maiztasunetan azaldu den bezala, fitxategi bakoitzak Train eta Test atalak ditu barnean, baina hauek ere banatuko dira. Probak egitean, konparazioak soilik Train atalarekin egingo direlako.

4.2.2 Maiztasuna

Metodo honetan, corpus batean neurtu egiten da laburdura baten hedapenak zein maiztasunekin agertzen diren eta beste batean hedapenak maiztasun hauen arabera esleitzen dira. Adibidez, *HTA*-ren corpusean azaltzen diren *HTA* guztiei, *HTA*-k duen hedapenen artean portzentaia altuena duenaren hedapena esleituko zaie. Kasu honetan, **hipertensión_arterial**, maiztasunen fitxategian %100 delako bere ehunekoa, naiz eta, *Histerectomía total abdominal* eta *Hipertensión arterial diastólica* ere hedapen posibleak diren. Hurrengo kasuan ikusten den moduan:

hta:

hipertensión arterial =100%

Hipertensión arterial diastólica =0%

Histerectomía total abdominal = 0%

batura=176

Horrez gain, laburdura bakoitzaren corpusean azaltzen den laburdura bakoitzari, hedapen maizenaz gain, zein karakteretan hasi eta bukatzen den ere adierazi beharko zaio, hori erreferentzia bat izango delako, etiketatzaileek bueltatuko dituzten etiketekin konparatzeko.

Hurrengoa, sistema honek bueltatuko dituen fitxategien adibide bat da:

T1 hipertensión_arterial 373 376 HTA

T2 hipertensión_arterial 455 458 HTA

T3 hipertensión_arterial 503 506 HTA

Adibide honetan ikusten den bezala, lehenengo zutabean, identifikadore bat azaltzen da, zenbaki bat. Honek egia esan, ez du garrantzia handirik beste zutabeekin konparazioan. Bigarren zutabean berriz, 5.zutabean dagoen laburdurak dituen hedapenen artean maiztasun altuena duena izango da. Kasu honetan, **HTA** laburduraren hedapenen artean maiztasun altuena duena, *hipertensión_arterial* da.

Horrez gain, hirugarren eta laugarren zutabeetan ateratzen diren zenbakiek, laburduren testu-fitxategietan laburdura zein karakteretan hasi eta bukatzen den esan nahi dute.

Bueltatzen diren fitxategiak, *.ann* erako fitxategiak dira (BRAT tresnak emaitzak formatu honetan ematen ditu). Horrez gain, 2 fitxategi aterako dira sistema hau erabiliz: A etiketatzailea eta B etiketatzailearekin konparatzeko emaitzak. Test-erako berriz, Train (probak egitean) ataletik aterako dira maizenak diren hedapenak eta Test-en probatuko dira. Honela, emaitzak hobetu beharko lirateke.

4.2.3 UKB (Grafoetan oinarritutako desanbiguazioa)

UKB (*Graph Based Word Sense Disambiguation and Similarity*) da erabiliko den hirugarren eta azken sistema. Aplikazio hau erabili ahal izateko, hiru elementu garrantzitsu behar dira. Eta elementu horiek ere diseinatu egin behar dira.

3 elementu horiek hiztegi bat, grafoa eta corpus-aren testuingurua dira. Ondorengo puntuetan, 3 elementu hauek azalduko dira.

Hiztegia

Atal honetan, hiztegiari buruz hitz egingo da. Hiztegian, termino-medikuak agertuko dira ondoan beraien UMLS (CUI - Concept Unique Identifier)-arekin. Horretarako, 3 fitxategi "iturri"erabili dira:

- **map_cui_sctspa.txt**

  UMLS (*Unified Medical Language System*)-en barnean, hainbat fitxategi daude, eta fitxategi horien artean, *Rich Release Format (RRF)* fitxategiak daude. RRF fitxategietako bat *MRREL.RRF* fitxategia da eta hemendik ateratzen da *map_cui_sctspa.txt* eta baita grafoa ere. *map_cui_sctspa.txt* SNOMED CT-ko kontzeptu identifikadoreen eta UMLS-ko CUI identifikadoreen arteko mapaketa duen fitxategia da.

  UMLSko taula guztiak datu-base erlazional batean kargatu egiten dira. Hau egin ondoren, *MRREL.RFF* fitxategiari, *mysql* galdera batzuk aplikatzen zazkio SNOMED CT - UMLS identifikadore pareak lortzeko. UMLS CUI-ak eta SNOMED CT identifikadoreak, hedapen berberari erreferentzia egin arren, ezberdinak dira, 2 hiztegi direlako (SNOMED CT, UMLSren barnean dago).

  map_cui_sctspa.txt fitxategiaren adibidea taulan ikusten da. Lehenengoan, terminoaren UMLS-ren kodeak daude (CUI-ak), bigarrenean be-

Taula 4.12: map_cui_sctspa.txt fitxategiaren adibidea

C0456965	419493001	6 metros *hallazgo*
C0052300	417889008	aceite de maní *sustancia*
C0003292	419028009	agente antidiarreico *sustancia*
C0018100	419163002	agente antigotoso *sustancia*
C0360390	419278002	agente antimicobacteriano *sustancia*
C0005515	418980009	agente biológico *sustancia*
	...	

rriz, terminoaren SNOMED CT-ren kontzeptu identifikadoreak eta hirugarrenean, terminoa.

UKB-rako erabiliko den hiztegian, fitxategiko 1. eta 3. zutabea erabiliko dira, baina hiztegian, terminoak 1.zutabean joango dira eta CUI-ak berriz, 2.zutabean.

Hipertensión arterial C0856216

...

Histerectomía total abdominal C0006512

...

hta C0856216 C0006512

Horrez gain, terminoekin batera "()"badaude, hauek kendu egingo dira. Adibidez, "6 metros *hallazgo*"beharrean, "6 metros"idatziko da.

- **FreelingFitxategiak.txt**

  Fitxategi hau, laburdura ez-anbiguoen fitxategia da, hau da, fitxategi honetan azaltzen diren laburdura guztiek hedapen bakarra dute. Fitxategi honetan, 3 zutabe bereizten dira baina, hiztegia sortzeko garrantzia, lehenengo bi zutabeek dute, lehenengoan laburdura eta bigarrenean honen hedapena agertzen delako. Badago, kasuren bat ere, hedapenik ez duena, eta laburdura gisa agertzen dena, hedapen moduan. Adib:

  vc2 vaso_concepción_2 NCOOLSO

  pga_2 pga_2 NCOOLSO

  mmm mielofibrosis_con_metaplasia_mietoide NCOOYLO

Fitxategi honen erabilera honako hau izango da: Laburdura (vc2) izanik, bere hedapena (vaso_concepción_2) hartuko da eta hau *map_cui_sctspa.txt* fitxategian bilatuko da bere **CUI**-ak zein diren eta hiztegian, laburdura eta honek dituen CUI-ei dagokien hedapenak esleituko dira. Hala ere, badira kasu batzuk, laburduraren hedapena ez dena agertzen *map_cui_sctspa.txt* fitxategian. Kasu horietan, laburdura horren hedapenaren CUI-a ez dagoenez, ez da hiztegian idazten.

- **LaburduraAnbiguoak_UTF8**

  Aurreko fitxategian, laburdura ez-anbiguoak agertzen baziren, fitxategi honetan berriz, laburdura anbiguoak azaltzen dira. Hau da, laburdura eta bere hedapen posibleak azaltzen dira. Honako moduan:

  ab.: ablación NCOOYLO // aborto NCOOYLO

  abd.: abdomen NCOOYLO // abducción NCOOYLO

  abv: adriamicina, bleomicina, y vinblastina, quimioterapia NCOOYLO // adriamicina, bleomicina, y vincristina, quimioterapia NCOOYLO

  Kasu hauetan, soilik bi hedapen posible daude, baina fitxategian hedapen gehiago dituzten fitxategiak ere azaltzen dira. Anbiguoak ez diren laburdurekin egin den gisa, hedapen bakoitzeko, bere CUI-a bilatuko dugu *map_cui_sctspa.txt* fitxategian eta hiztegian idatziko dugu. Beraz, kasu honetan, hiztegian, laburdurak hainbat CUI izango ditu:

  hiztegia.txt :

  tm C0002875 C0026565 C0198161 C0027651

  tpl C0442951

  tp C0033707 C0040034 C0041327

  Kasu hauetan, *tm*-k 4 hedapen posible ditu eta hauen CUI-a du alboan, bai *tpl* eta *tp*-k bezala. Etiketatzaileek eskuz etiketatu dituzten fitxategietan, analisi bat egingo da, zein hedapen erabiltzen dituzten ikusteko eta hedapen bat fitxategi honetan ez dagoena ugari erabiltzen badute, hedapen hori fitxategi honetan txertatuko da.

Laburdurak, hiztegian modu horretan agertuko dira, eta CUI identifikadore horiek, beraien hedapenek ere izango dituzte.

Grafoa

UKB erabiltzean, beharrezkoa den beste baliabide bat, grafoa da. Grafoa, erpin eta erlazio bidez dago osatua. Erpinak, kontzeptuak dira eta erlazioak, kontzeptuen arteko erlazioa. Grafoaren erabilera laburduren desanbiguazioan honako hau da: Laburdura bat hartzean, testuinguru kontzeptuak hartuz gertuen dituen kontzeptuen artean, gertuen agertzen den hedapena izango da aukeratua.

Beraz, UKB-k grafoa erabiliko du, testuinguru eta hiztegia erabiliz, grafoan dauden kontzeptuen (hedapenen) artean gertuena bilatzeko.

Gainera, grafoa, *MRREL* taulatik sortzen da. Taula hau, *UMLS Full Release File*-n agertzen da, beraz, UMLS-ren webgunean lortu ahal daiteke. Memoria karga handia duen taula da MRREL fitxategia, eta bertan UMLS-ko kontzeptuak daude eta horien bidez osatzen da grafoa. Kontuan izan behar da, UMLS-k SNOMED CT duela barneratua. Beraz, proiektuan, SNOMED CT behar dugu (gazteleraz delako), baina, UMLS guztiaren grafoa izango dugu.

Testuingurua

UKB-k laburdura anbiguo bat hartzean, bere inguruan dauden hitzak kontuan hartzen ditu, grafoan hitz horietaz baliatuz, hedapen egokiena aukeratzeko. Beraz, corpusa beharrezkoa da UKB-n, baina ez corpus-a dugun gisa. Corpus-eko esaldi bakoitzeko hitzei honelako formatua ematen zaie:

no##t31#0#1

"no"forma da, eta t31 hori, corpus-ari *Freeling-med* aplikatu eta gero sortzen den **KAF** fitxategain "no-ri ematen zaion "t"identifikadorea da. Horrez gain, "0-ak esan nahi du hitza ez dela anbiguoa. Anbiguoa izango balitz, "1"jarriko litzateke. Azkenik, azkeneko zenbakia "pisua-ri deritzo. Proiektu honetan, denei pisu bera emango zaie, "1".

Hona hemen, *HTA* eta *DM* laburdurak dituen esaldi bat:

no##t31#0#1 hta##t32#1#1 no##t33#0#1
dm##t34#1#1

Esaldi bakoitzak lerro bat izango du, eta gainera, ikusten den bezala, puntuazioak ez dira agertzen.

Hau guztia azaldurik, gauza garrantzitsu bat azaldu behar da: Hitzkonposatuak direnean, espazioak "_"batekin joan behar dute, adibidez, "volumen_pulmonar".

SNOMED CT-ko terminoak hiztegian "_-rekin lotuta agertu behar dute eta testuinguruan ere bai. Bestela, ez lukete bat egingo.

Kapitulua 5

Inplementazioa

Atal honetan, proiektuan zehar osatu diren programen inguruan hitz egingo da. Programa bakoitzari dagokion informazio orokorra emango da.

Proiektuan zehar egin diren programak, *Perl* lengoaian idatziak izan dira, erraztasunak ematen dituelako lengoaia naturalaren inguruan erabiltzen diren testu eta adierazpen erregularrekin lan egiteko.

5.1 Corpusaren eraketa

Honakoak dira corpusa eratzeko erabili diren programak:

5.1.1 Azpicorpusa sortu

- **SARRERA:** Direktorio bat. Direktorio honetan, hainbat fitxategi daude. Fitxategi bakoitza hainbat esaldiz dago osatua.
- **IRTEERA:** Fitxategi bat: *Corpusa.txt*.

Programa honen bidez, corpus bat sortzen da. Horretarako, sarrera gisa sartzen den direktorioko fitxategiak hartu eta hauetatik, 5 letra baino gutxiagoko hitzen bat duten esaldiak gorde eta, ausaz esaldi kopuru batzuk idatziko dira irteerako fitxategian.

Modu honetan, hasiera batean erabiliko den corpus-a sortzen da. Programa honen helburua, baimenduta ez dauden pertsonak, konfidentzialak ez diren

fitxategiak ez ikustea da. Honela, fitxategi horietatik "x"esaldi hartuko dira, eta ez da fitxategia irakurriko.

5.1.2 Corpus-iturria zatika banatu

- **SARRERA:** Corpus-iturria. Galdakao 2012-ko txostena.
- **IRTEERA:** Direktorio bat.

Corpu-iturria zatika banatzen du, paragrafo bakoitza fitxategi batean gordez.

5.1.3 Corpus definitiboa sortu

- **SARRERA:** Aurreko programak bueltatzen duen direktorioa.
- **IRTEERA:** Bi fitxategi: *Emaitza.txt* eta *laburdurak.txt*.

Programak, direktorioko fitxategiak hartu eta, hauetan, esaldika, 5 letra baina gutxiagoa den hitzak badaude ("stop-words kenduta"), esaldi hauek, *emaitza.txt* fitxategian gordetzen joango da. Esaldiekin batera, fitxategi bakoitzeko esaldien batazbestekoa zein den ere azalduko da.

laburdurak.txt-n berriz, sarrera direktorioko fitxategi guztietan agertzen diren laburdura guztiak idatziko dira, beraien agerpen kopuruarekin. Gainera, laburdurak ordenatuta joango dira, agerpen gutxien dituztenetatik, agerpen gehien dituztenetara.

5.1.4 Zutabeka idatzi laburduren ezaugarriak

- **SARRERA:** *emaitzak.txt* eta *laburdurak.txt*.
- **IRTEERA:** Fitxategi bat : *zutabeka.txt*.

Laburdura bakoitzeko, honakoa idatziko da:

Laburdura + Esaldien luzera + Esaldia + Fitxategia

Adibidez:

hspace3cm*DM 2 No DM fitxategi-1*

5.1.5 Laburdura bakoitzeko corpusa sortu

- **SARRERA:** Fitxategi bat: *ZUTABEKA.txt*.
- **IRTEERA:** Direktorio bat: *laburdurak*. Bertan, aukeratutako laburdurako direktorio bat egongo da eta direktorio horietako bakoitzeko, 3 fitxategi sortuko dira: *denak.txt*, *txiki.txt* eta *handi.txt*.

Programa honek, *ZUTABEKA.txt* fitxategia hartzen du, eta aukeratu diren laburdura bakoitzeko, bere esaldi luzeraren arabera, gordeko dira fitxategi zehatz batzuetan. Hau da, laburduraren esaldia, 5 edo hitz kopuru gutxiagokoa bada, *txiki.txt* fitxategian gordeko da eta baita *denak.txt* fitxategian. Bestalde, laburdura agertzen den esaldiaren hitz kopurua 5 baina handiagoa bada, *handi.txt* fitxategian eta *denak.txt* fitxategina gordeko da.

Hau da, *denak.txt* fitxategian, aukeratutako laburduraren agerpen esaldiak idatziko dira. *txiki.txt*-n berriz, 5 letra edo txikiagoak diren esaldiak eta *handi.txt*-n 5 letra baina handiagoak direnak.

5.1.6 Esaldi txikiak ondo egituratu

- **SARRERA:** Direktorio bat : *laburdurak*.
- **IRTEERA:** *laburdurak* direktorioan dagoen direktorio bakoitzean (laburduren izenak) fitxategi berri bat bueltatuko du: *txiki-definitiboak.txt*.

banatu-laburdura.pl programa exekutatzean, *txikiak.txt* fitxategia sortzen zen beste bi fitxategiekin batera. Baina, fitxategi honetan, esaldiak 5 edo hitz gutxiagoekin osatutako esaldiak sartzen dira. Honek esan nahi du, "*No DM*"edo "*No HTA*"bezalako esaldiak fitxategi honen barruan sartzen direla. Honek, ez du testuingururik eta beraz, UKB erabiltzerakoan, posibleena da datu txarrak ematea da.

Hori saihesteko, 3 hitz baina gutxiago dituzten esaldiak baztertuko dira, eta horretarako, programa honek, fitxategi berri bat sortuko du, proiektuan zehar erabiliko diren esaldi "txikiak"sartzen direnak, hau da, 3 eta 5 arteko hitz kopuruak dituzten esaldiak.

5.1.7 Etiketatzaileentzat corpusa banatu etiketatzeko

- **SARRERA:** Fitxategi bat: *TAULA.txt*(Taula 3.3). Fitxategi honetan, laburdura bakoitzeko, *handiak.txt*, *txikiak.txt* eta *denak.txt* fitxa-

tegiei, terminal batean **wc -l fitxategia** aplikatu ondoren, duten esaldi kopurua agertzen da.

- **IRTEERA:** *laburdurak* direktorioan dauden laburdura bakoitzeko direktorioan, 2 fitxategi berri sortuko dira: *A.txt* eta *B.txt*.

Programa honek, bi etiketatzaileei banatuko zaizkien testuak osatzen ditu. *3.1.4* atalean azaltzen den moduan, laburdura bakoitzeko 900 esaldi banatuko dira. Programa honek, laburdura bakoitzeko esaldi txiki eta handiak modu proportzionatuan banatzen dizkie bi etiketatzaileei. Esaldi bai handi eta txikiak etiketatzaile bakoitzari proportzio berdinean banatzen zaizkie.

Etiketatzaile bakoitzari 525 esaldi banatuko zaizkie, horietatik bakoitzaren lehenengo 375 esaldiak desberdinak izango dira, baina, fitxategietan, bai A.txt eta B.txt fitxategian, amaieran dauden 150 esaldiak komunak izango dira bi etiketatzaileentzat.

5.1.8 Corpus-a Train eta Test-en banatu

- **SARRERA:** Direktorio bat: *laburdurak*.
- **IRTEERA:** *laburdurak* direktorioan dauden direktorio bakoitzean 2 fitxategi: *train.txt* eta *test.txt*.

Programa honek, *laburdurak* direktorioan , laburdura bakoitzeko direktorioan, *train.txt* eta *test.txt* fitxategiak sortzen ditu. *train.txt* fitxategian, entrentatzeko erabiltzen diren fitxategiak gordeko dira, eta *test.txt* fitxategian berriz, azken emaitzetarako erabiliko diren esaldiak.

5.1.9 Etiketatzaileei banatutako corpusa banatu Train eta Test-erako

- **SARRERA:** Direktorio bat: *laburdurak*.
- **IRTEERA:** *laburdurak* direktorioan dauden laburdura direktorio bakoitzeko 6 fitxategi: *A-train.txt*, *A-test.txt*, *B-train.txt*, *B-test.txt*, *eq-train.txt* eta *eq-test.txt*.

Programa honek, laburdura bakoitzeko, etiketatzaile bakoitzari banatu zaizkion fitxategiak train eta test-erako banatuko dira. Hau da, A.txt-etik, *A-train.txt*, *A-test.txt*, *eq-train.txt* eta *eq-test.txt* aterako dira eta B.txt-etik *B-train.txt* eta *B-test.txt*.

3.1.5 atalean azaltzen den gisa, etiketatzaile bakoitzak dituen testuetatik batzuk probetarako eta beste batzuk azken emaitzetarako erabiliko dira.

5.1.10 Etiketatzaileek komunak dituzten fitxategiak Train eta Test-erako banatu

- **SARRERA:** Laburdura bakoitzeko direktorioan dagoen *eq-test.txt* eta *eq-train.txt* fitxategiak.
- **IRTEERA:** *laburdurak* direktorioaren barruan dauden direktorio bakoitzaren barnean bi fitxategi: *eq-test.ann* eta *eq-train.ann*.

Programa honek, sarrera moduan sartzen zaizkion fitxategi bakoitzarekin, fitxategi bakoitzean, subprograma baten bidez, laburdura bat aurkitzean, .ann fitxategi batean idatziko da zein karakteretan hasi eta bukatzen den laburdura. Honela, train fitxategia sarrera gisa denean, .ann train-ena izango da eta test-ekin egitean, .ann test-ena.

Irteera moduan ematen diren bi fitxategi hauek, train-en azkeneko hedapen agerpena noiz ematen den jakiteko da. Etiketatzaileek, *A_eq.ann* eta *B_eq.ann* bietan train eta test nahastuta eman dituztelako, programa honetako fitxategiak, erabiliko dira *A_eq.ann* eta *B_eq.ann* fitxategiak train eta test-en banatzeko.

5.2 UKB-rentzako prestaketa

5.2.1 KAF fitxategiak prozesatu

- **SARRERA:** Freeling-med corpusari aplikatu ondoren sortzen den *KAF* fitxategia.
- **IRTEERA:** *tmp* izeneko direktorio bat. Bertan, *.term* fitxategiak izango dira.

Programa honek, sarrera bezala, fitxategi bateri Freeling-med aplikatzean sortzen den *KAF* fitxategia pasatzen zaio eta hau izanik, .term fitxategiak sortzen ditu hurrengo formatuarekin:

t3 81 81 Z

t4 aos ao NCMP000 calificador

t5 VARON varon VMIS3P0

t6,0 MC masa_corporal NC00YL0

t6,3 MC motivo_de_consulta NC00YL0 hallazgo

t6,2 MC mdico_de_cabecera NC00YL0 ocupacin

t6,1 MC media_cuadrtica NC00YL0

5.3 UKB-rako hiztegia sortu

- **SARRERA:** 3 textu-fitxategi:*map_cui_sctspa.txt*, *FreelingFitxategiak.txt* eta *LaburduraAnbiguoak_UTF8*.
- **IRTEERA:** Testu-fitxategi bat: *hiztegia.txt*.

Programa honen bidez, *map_cui_sctspa.txt* fitxategia garbituko da eta gainera, beste bi fitxategietako laburdurak beraien hedapenen CUI-ekin, *hiztegia.txt*-n gordeko dira.

map_cui_sctspa.txt garbitu egingo da, fitxategiak, hitz askotarako, *(substancia)* edo *(concepto)* agertzen badira adibidez, ezabatu egin beharko dira.

5.3.1 KAF fitxategietatik, testuingurua sortu

- **SARRERA:** *tmp/.terms* izango dira(banaka). Fitxategi hauek, *Freeling-med* corpusari aplikatu ondoren sortzen diren *KAF* fitxategiei *BowAtera.pl* programa aplikatzean sortzen diren fitxategiak dira.
- **IRTEERA:** *kontextu* izeneko direktorio bat sortuko da, non bertan, sarrera gisa sartzen den *.term* fitxategiari, bere *.txt* fitxategia sortuko zaion.

Programa honek, UKB erabiltzean behar den 3 parametroetatik *testuingurua* deritzoguna sortzeko balio du. Irteera gisa ateratzen den textu-fitxategian, esaldi bakoitzeko, hitz bakoitzeko hainbat datu ezartzen dira diseinuaren atalean azaldu den bezala.

5.3.2 UKB - exekutatu

UKB exekutatzeko, terminalean, bi komando exekutatu behar dira:

- *./compile_kb -o kb_umlsmrrel.bin kb_umlsmrrel.txt*
- *./ukb_wsd –ppr_w2w –nopos -K kb_umlsmrrel.bin -D hiztegia testuinguruа*

Lehenengo komandoan, grafoa .txt fitxategi batetik, .bin fitxategira pasatzen du, bigarren komandoan erabili ahal izateko. Bigarren komandoan, sarrera bezala, lehenengo komandoak emandako grafoa, hiztegia eta testuinguruaren fitxategia sartzen zaizkio.

Hau exekutatu ondoren, terminalean bertan, UKB-k bere emaitzak ematen ditu, laburdura bateko corpusean, laburdura hori ez diren beste laburdurak ere anbiguatzen ditu.

Emaitzen formatua honakoa da:

ecg##t1#1#1 t14 C1623258 !! ecg zona##t29#0#1 t32 C1623258 !! ecg ecg##t50#1#1 t79 C1623258 !! ecg capacidad_funcional##t92#0#1 t94 C1623258 !! ecg no##t107#0#1 t113 C1623258 !! ecg

Azkeneko zutabean, desanbiguatu duen laburdura agertzen da, hirugarren zutabean berriz, laburdurari esleitu dion hedapenaren CUI identifikadorea eta bigarren zutabean, testuinguruan laburdurak duen "t"identifikadorea. Modu horretan, jakiten da zein den desanbiguatu duen laburdura testuinguruan.

5.4 Maiztasunen sistema

5.5 Etiketatzaileen aurreko etikezioetatik, laburduren hedapenen maiztasunak lortu

- **SARRERA:** Fitxategi bat: *Miriam+SaraSortLatin1.txt*. Fitxategi honetan, 2 etiketatzaileek beste corpus bati egin dizkien etiketazioak agertzen dira.
- **IRTEERA:** Fitxategi bat: *maiztasunak.txt*.

Programa honek, 2 etiketatzaileen etiketazio-fitxategi bat hartu eta hemendik, fitxategian laburdura bat agertzen den bakoitzeko, zein hedapen posible dituen eta aukeratua zein den jasoko da. Azkenean, laburdura bakoitzeko,

zein hedapenez gain, hedapen bakoitza zenbat aldiz hedapen posible gisa aukeratua izan den jasoko da. Modu horretatik, laburdura bakoitzeko, hedapen bakoitza aukeratua izan den ehunekoa jasoko da.

maiztasunak.txt fitxategian, parametro gisa sartu den fitxategian agertzen den laburdura bakoitzeko, bere hedapenak, hauen portzentaia eta zenbat aldiz agertu den idatziko dira. Beraz, *maiztasunak.txt* fitxategian, beste corpus horrekiko maiztasunak aterako dira eta horiek erabiliko dira proiektuan, maiztasunen sisteman Trainerako.

5.5.1 Etiketatzaileekin konparatzeko .ann fitxategiak sortu

- **SARRERA:** Direktorio bat: *laburdurak*.
- **IRTEERA:** Laburduraren direktorio bakoitzeko, 4 fitxategi: *"Labizena-A-train.ann*, *"Labizena-A-test.ann*, *"Labizena-B-train.ann* eta *"Labizena-B-test.ann*

Programa honek, maiztasunen sisteman etiketatzaileen fitxategiekin konparatzeko erabiliko diren fitxategiak sortzen ditu. Hau da, A etiketatzaileak komunak ez diren esaldietatik trainerako direnak, *"Labizena-A-train.ann* fitxategiarekin konparatuko dira, eta B etiketatzailearen kasuan, *"Labizena--B-train.ann* fitxategiarekin.

Fitxategi hauek sortzeko, programak, laburdura bakoitzak dituen direktorioetatik, subprograma baten bidez, *etiketatzaileen-train-test-banatu.pl* programak bueltatzen dituen fitxategietan laburdura azaltzen den karakterea eta bukatzen den karaktere zenbakiak hartuko ditu eta ondoren, *.ann* fitxategi batean joango da idazten, laburduraren hasierako karakterea, bukaerako karakterea eta *maiztasunak.txt*-n laburduraren hedapen portzentaia altuena duen hedapena.

Hori egingo da, A eta B etiketatzaileei banatu zaizkien bai test eta train atalekin, beraz, lau *.ann* fitxategi osatuko dira.

5.6 Ausazkotasunen sistema

5.6.1 Etiketatzaileekin konparatzeko ausaz sortutako sistemaren emaitzen .ann fitxategia lortu

- **SARRERA:** Direktorio bat: *laburdurak*.
- **IRTEERA:** Laburduraren direktorio bakoitzeko, 4 fitxategi: *"Labizena-A-asusa-train.ann*, *"Labizena-A-ausa-test.ann*, *"Labizena-B-ausa-train.ann* eta *"Labizena-B-ausa-test.ann*

Programa honek *annotation.pl* programaren berdina egiten du, baina *maiztasunak.txt* fitxategian laburdura bakoitzeko hedapen portzentaia altuena duena aukeratu beharrean, programa honek, *maiztasunak.txt*-en laburdurak dituen hedapenen artean bat aukeratzen du ausaz, eta hedapena agertzen den bakoitzean, *maiztasunak.txt*-n laburdurak dituen hedapenetatik bat ausaz aukeratuko da. Beraz, *annotation.pl*-k laburdura agertzen zen kasu bakoitzean, hedapen berdina jartzen zuen: maiztasun handiena zuenarena, eta programa honetan berriz, laburdura agertzen den bakoitzeko, hedapen bat ausaz aukeratuko da. Ez da kasu guztietan hedapen bera egongo.

5.7 Etiketatzaileekiko konparazioak

5.7.1 Interannotator-agreement

- **SARRERA:** Direktorio bat: *laburdurak*. Direktorio honetan, etiketatzaileek etiketatutako fitxategiak ere gorde behar dira, aurretik zeuden fitxategietaz gain.
- **IRTEERA:** Direktorio bat: *interannotator*. Direktorio honetan, proiektuan zehar probetan erabiliko diren laburduren izenak dituzten direktorioak egongo dira. Direktorio bakoitzean programak 9 fitxategi bueltatuko ditu: *BAI-A.txt*, *BAI-B.txt*, *BAI-eq.txt*, *EZ-A.txt*, *EZ-B.txt*, *EZ-eq.txt*, *OTRO-A.txt*, *OTRO-B.txt* eta *OTRO-eq.txt*,

Programa honek, etiketatzaileek etiketatutako etiketatzioak eta gure sistemak emandako fitxategiak(maiztasunen eta ausaz bidezkoak) konparatzen ditu. Konparazioan, *BAI* fitxategiek, hedapenaren emaitzean berdinak direnak kontatzen ditu, *EZ*-ek berriz, desberdinak direnak, eta *OTRO* kasua, etiketatzaileek "otro"jartzen duten kasuetarako ematen da.

Kapitulua 6

Emaitzak

Atal honetan, proiektuan eman diren sistemak etiketatzaileen etiketazioaren aurka izan dituzten emaitzak azalduko dira. Hiru ataletan banatuko da: Ausazkotasunen emaitzak, maiztasunen emaitzak eta UKB-ren emaitzak.

Esan beharra dago, hasiera batean, paragrafo bakoitzeko laburdura bat zegoela pentsatzen zen, eta hori ez da horrela, beraz, kasu batzuetan laburdura kopurua handiagoa da. Bestalde, etiketatzaileei, hainbat zehaztapen jarri zitzaizkien etiketaziorako. Hau da, etiketatutako laburdurak, soilik dijoaztenak etiketatu behar zituzten, hurrengo moduan:

No HTA: **EZ**

No HTA no DM **BAI**

Hau da, puntuazioak badituzte bai aurrean edo atzean, laburdura hori ez da etiketatuko. Etiketatzaileek kasu askoretan, laburdura mota horiek ere etiketatu dituzten, etiketatu behar ez zirenean.

Gainera, sistema bakoitzaren emaitzak ematean, hauen estaldura eta doitasuna azalduko dira. Estaldura, laburdura guztietatik, asmatze-ehunekoa da eta doitasuna berriz, etiketatu direnetatik ateratako asmatze-tasa. Hau da, 100 esaldi badira etiketatu behar direnak eta soilik 80 etiketatu dira. Gainera, etiketatu diren 80etatik, 60-etan asmatu du. Beraz,

Estaldura : $60/100 = \%60$.

Doitasuna : $60/80 = \%75$.

6.1 Laburdura kopurua corpusean

Hurrengo taulan, fitxategi bakoitzeko laburdura kopurua azalduko da eta datu hauetatik lortuko da estaldura bai, ausazkotasunean oinarriutako , maiztasunean oinarritutako eta UKB sistemena.

Taula 6.1: Laburdura kopurua fitxategi bakoitzeko

Laburdurak	**A-train**	**A-test**	**B-train**	**B-test**	**Train**	**Test**
C	277	72	280	68	557	140
CD	330	85	347	78	677	163
DA	358	93	367	96	725	189
DM	261	62	256	60	517	122
EAC	287	73	298	74	585	147
ECG	296	74	301	75	597	149
FA	301	75	301	74	602	149
FC	305	74	307	75	612	149
h	272	64	265	68	537	132
HTA	249	62	243	64	492	126
K	302	75	300	76	602	151
PCR	299	74	294	76	593	150
QRS	320	78	331	77	351	155
T	335	87	324	81	659	168
TA	296	75	300	73	596	148
TAC	299	77	305	80	604	157
TP	309	76	306	78	615	154
VI	297	75	299	75	596	150

Sistemek emango dituzten emaitzen estaldura, taula honetako datuekin konparazioan aterako da. Hau da, *Laburdura kopurua fitxategi bakoitzeko* taulako datuak izango dira erreferentzia. Esan beharra dago, ikusten den moduan, hasiera batean eman ziren esaldi kopuruekin guztiz bat ez datozela, arrazoi batengatik. Etiketatzaileei etiketatzeko arau gisa, puntuazio artean zeuden laburdurak ez etiketatzeko esan zitzaien, beraz, orain, datuak lortzerakoan, erregela hori jarraitu da, eta kasuren batzuetan, laburdura kopurua jaitsi egin da. Adibidez, *C*-ren kasuan, Train osoa 600 esaldikoa zen baina, 557 laburdura erabiliko dira horietatik probetarako.

Bestalde, badira kasuak, non esaldi batek hainbat laburdura zituen, eta kasu horietan, laburdura kopurua, esaldi kopurua baina handiagoa da. Adibidez,

TAC-en kasuan, Train-en, 600 esaldietatik, 604 laburdura daude. Datu guzti hauek erabiliko dira, sistemek ematen dituzten emaitzen estaldura kalkulatu ahal izateko.

Datu hauek izanda, 3 sistemetan estaldura zenbatekoa jakin behar da. Estaldura esaterakoan, zenbat laburdura detektatu eta desanbiguatu dituen esan nahi du.

Horretarako, sistema bakoitzeko, estaldura eta doitasuna azalduko dira hurrengo ataletan:

6.2 Ausaz-baseline

Ausazkotasunetan oinarritutako sistemak eman dituen emaitzetan, fitxategi bakoitzeko laburdura kopurua, aurreko atalean eman direnen bera da. Hau da, sistemak, laburdura guztiak etiketatu ditu. Beraz, kasu horretan, doitasuna eta estaldura berak dira.

Taula 6.2: Ausazkotasunen doitasuna eta estaldura

Laburdurak	**Train**	**Test**
C	%0	%0
CD	%32,84	%38,12
DA	%50,32	%40,82
DM	%99	%100
EAC	%43,35	%52,83
ECG	%99,1	%99,85
FA	%33,62	%32,22
FC	%47,85	%51,83
h	%32	%35,62
HTA	%50,38	%46,8
K	%29,42	%30,45
PCR	%29,1	%35,83
QRS	%49,31	%47,8
T	%1,702	%2,65
TA	%47,8	%42,01
TAC	%48,295	%51,15
TP	%29,245	%35,02
VI	%94,6	%93,3

Ausas oinarrituta dagoen sisteman, emaitzak, logikoki eskasak dira, ausaz harztzen direlako hedapenak eta beraz, orokorrean logikogoa da emaitzak txarragoak izatea. C eta T-ren kasuak oso txarrak dira, lehenengoaren kasuan estaldura eta doitasuna 0-koa da. Honen arrazoia, etiketatzaileek "otro"bezala etiketatu dituztela etiketak.

Ikusi daiteke ere, laburdura tauletan (ikus 4.1-4.10 taulak), aukeraketa egitean, hedapen bakarra zuten kasuan, portzentaiak ez direla %100koak, hori etiketatzaileek kasu guztiak ez etiketatzearen arrazoia delako.

Kasu hauetan, train eta test-en arteko doitasuna ez da oso desberdina, bi kasuetan, hedapen berdinak ausa moduan erabiltzen direlako jakiteko zein den hedapen aukeratua.

6.3 Maiztasunak

Ausazkotasunean oinarritutako baseline sisteman moduan, kasu hauetan ere, sistemak fitxategiko laburdura guztiak etiketatu ditu eta beraz, doitasuna eta estaldura berdinak dira.

Taula 6.3: Maiztasunen doitasuna eta estaldura

Laburdurak	**Train**	**Test**
C	0%	0%
CD	%99,42	%99,35
DA	%96,87	%79,67
DM	%99,41	%100
EAC	%95,76	%97,26
ECG	%97,51	%97,32
FA	%46,95	%59,02
FC	%96,88	%99,32
h	%99,24	%99,30
HTA	%94,69	%95,21
K	%95,66	%98,66
PCR	%89,70	%93,35
QRS	%97,95	%98,71
T	%4,72	%5,74
TA	%97,66	%97,96
TAC	%99,12	%98,22
TP	%72,05	%71,185
VI	%92,45	%93,4

Maiztasun bidezko doitasuna taulan ikusten den moduan, emaitzak orokorrean, oso onak direla esan behar da. Ia kasu guztietan ehunekoa oso altua da, C eta T laburduren kasuetan izan ezik. Arrazoia, etiketatatzaileek, laburdurak kasu horietan "otro"gisa etiketatzearena da.

6.4 UKB

UKB-ren kasuan, ezberdina da. UKB-k ez ditu beti laburdura guztiak desanbiguatzen eta beraz, estaldura oso ezberdina da beste bi kasuekin konparatuz. Kasu batzuetan, bai *h* , *DM* ,*K* eta *T* laburduren kasuan, UKB-k ez du ezer ere desanbiguatzen, *Freeling-med* aplikatzen zaienean testu-fitxategiei, honek, hirurak desanbiguatzen dituelako bere testuinguruaren arabera. Beraz, UKB-k ez ditu laburdura anbiguo gisa hartzen, Freelin-med-ek desanbiguatu dituelako

Adibidez, *DM*-ren kasuan, testuingurua beti *DM tipo II* denez, *Freeling-med* berak, zuzenean desanbiguatzen du *Diabetes Tipo II* gisa. Beraz, esan beharra dago, 4 laburdura hauek, erabili den corpusean ez direla laburdura anbiguoak.

Esan beharra dago, laburdura bakoitzeko UKB-k prozesatzen duen fitxategian, laburduraz gain beste hainbat laburdura ere badaudela, eta UKB-k hauek ere desanbiguatzen dituela, kontuan ez eduki arren proiektu honetan.

Taula 6.4: UKB estaldura

Laburdurak	**Train**	**Test**
C	%0	%0
CD	%5,61	%54,60
DA	%0	%100
DM	-	-
EAC	%80,34	%48,29
ECG	%78,57	%95,97
FA	%66,27	%6,04
FC	%43,79	%78,65
h	-	-
HTA	%54,47	%95,20
K	-	-
PCR	%0,67	%39,73
QRS	%0	%0
T	-	-
TA	%70,30	%44,59
TAC	%48,34	%30,57
TP	%0	%0
VI	%50,50	%47,33

Estaldurak ikusirik, doitasuna ateratzen da. Doitasuna, lortu diren emaitzen konparazioan datza, hau da, laburdura ugarietan, ez ditu laburdura agerpen guztiak desanbiguatu. Baina kasu horietan, desanbiguatu direnen artean, asmatze-tasa kalkulatuko da. *UKB teknikaren doitasuna* taulan ikus daiteke *UKB*-ren doitasun taula.

Taula 6.5: UKB bidezko doitasuna

Laburdurak	**Train**	**Test**
C	%0	%0
CD	%20,20	%93,68
DA	%0	%99,08
DM	-	-
EAC	%100	%100
ECG	%100	%100
FA	%78,96	%81,81
FC	%100	%100
h	-	-
HTA	%100	%100
K	-	-
PCR	%50	%92,30
QRS	%0	%0
T	-	-
TA	%98,59	%100
TAC	%100	%100
TP	%0	%0
VI	%95,28	%94,66

Tauletan ikusten den moduan, UKB-k ez ditu maiztasunak baino emaitza hobeagoak ematen orokorrean. Kasu batzuetan, emaitzak oso onak dira, adibidez, *TA* , *TAC* , *HTA* , *ECG...* non test-ak hartuz, hauen doitasuna %100ekoa den.

Hala ere, beste hainbat kasu daude, estaldura 0-koa ez izanik doitasuna bai den 0koa. Horrek pentsarazten du, arazoa, UKB-k emaitza egokiak ez emateko, hiztegian egon daitekeela, non hedapen guztien CUIak ez diren agertzen. Horrek bere eragina izan dezake, grafoan testuinguruan dauden hitzen CUIen bidez hedapen egokiaren bila ibiltzean.

Adibidez, *QRS* laburduraren kasuan, UKB-k *quiste renal simple* ematen du laburdura guztien hedapen gisa, baina etiketatzaileentzat (baita maiztasunen sisteman), bere hedapen posiblea *Parte del trazado del electrocardiograma que representa la despolarizacion ventricular* da. Hiztegian, azken hedapen hau ez dago, beraz, esan daiteke, posibleenik, horrek, UKB-k emaitza gaizki bueltatzea eragin duela.

Bestalde, *TP*-ren kasuan, laburduren tauletan (ikus 4.4-4.10 taulak) aukeratu ziren laburdurak hiztegian badaude, baina UKB-k agerpen guztiei, *trombopenia* hedapena esleitzen die hedapen posible gisa. Hedapen hau ez da inoiz agertzen etiketatzaileen etiketazioan.

Kapitulua 7

Ondorioak eta etorkizunerako lanak

7.1 Ondorioak

Behin proiektua bukatutzat emanda, honen inguruan gertaturiko gora-behera guztiak analizatzeko momentua da. Jarraian proiektuak izan dituen ondorioak zein ondorio pertsonalak deskribatzen dira.

7.1.1 Proiektuaren ondorioak

Proiektuan zehar osatzen joan diren galderen hainbat ekarpen eman dira proiektua amaitzean.

Labudura ez-anbiguoak corpus honetan

Alde batetik, argi ikusi ahal izan da emaitzetan, hainbat laburdura ez direla anbiguoak erabili den corpus-ean. Teorian, laburdura horiek anbiguoak dira, baina testuaren arabera ez dute zertan anbiguo izan, hau da, corpus batean laburdura bat agertzen den agerpen guztietan hedapen bera badu, ez litzateke zuzena corpus horretan laburdura hori anbiguo gisa onartzea. Hau da hain zuzen ere, GAP honetan gerturatu dena. Anbiguo uste genituen hainbat laburdura, modu bakarrean erabiltzen dira osasun-txosten hauetan.

Honelako hainbat adibide izan ditugu proiektuaren emaitzetan. Alde batetik, eskuzko etiketatzearen ondorioz hedapen bakarrak zutela konturatu ginenak (ikus 4.4-4.10 taulak, adibidez DM): DM eta ECG kasuak. VI ere, beti *ventriculo izquierdo* zen baina, etiketatzaileek eskuz, kasu batzuetan "otro"gisa etiketatu zuten "Seis"adierazteko, beraz, taulen aukeraketan, "otro"ere aukeratu zen.

Beste batetik, oso argiak direnak, *Freeling-med* berak desanbiguatu ditu testuinguruaren arabera eta honi hedapen posible bakarra eman. Kasu horien artean, *DM* (tauletan, anbiguo ez zela aukeratu zen) , *K*, *h* edo *T* laburdurak daud., Hauek *Freeling-med* aplikatzean, desanbiguatu egiten dira eta horrez gain, bueltatzen den KAF fitxategian, ez-anbiguo gisa bueltatzen ditu zuzenean.

Arrazoia, SNOMED CT-n beste kontzeptu konplexuago baten barruan azaltzen duelako. Adibidez, 'diabetes mellitus' beti agertzen da 'diabetes mellitus tipo II' barruan. Tokenizatzaileak bildu egiten ditu.

Kasu horietaz at, badira erabilitako corpusean, *Freeling-med*-ek corpusa analizatzean, anbiguoak diren laburdurak eta *UKB*-k desanbiguatzean, kasu guztietan hedapen bera dutenak. Hau da, anbiguo izan arren, corpus honetako kasu guztietan hedapen bera dute. Kasu horien artean, adibidez, *HTA* laburdura dago, corpus-aren kasu guztietan, *Hipertensin arterial* hedapena duena.

Beste ondorioen artean, *UKB*-ren erabilera dago. *UKB* hasiera batean, esaldi motzekin eta beste corpus batekin probatu zen, honen emaitzak onak ziren ikusi ahal izateko. Emaitzak, oso kaxkarrak izan ziren eta corpus-aren kasuan emaitzak desberdinak dira, hau da, testuinguru egokia eta zabala denean, *UKB*-k emaitzak hobetzen ditu.

Hala ere, UKB-ren emaitzak ez dira maiztasunetan oinarritutako sistemaren emaitzak baino hobeak. Gehienbat, kasu batzuetan, *UKB*-k ez du ezta laburdura bat ere desanbiguatzen eta beraz, kasu horietan estaldura 0 da.

UKB-k hedapen posible guztiak ondo ez ematea ez da normala, UKB grafotan oinarrituta egonik, eta testuingurua emanik, emaitza hobeak eman beharko lituzkeelako. Horren arrazoi bat honako hau izan daiteke: *UKB*-k hiztegi bat du bertan, hedapenen eta laburduren CUI-ekin. Kasuren batean, posible izan da, *SEDOM*-n laburdurak dituen hedapen posible guztiak ez egotea hiztegian, eta kasu horretan eragina sortu ahal izan du azken emaitzean.

UKB-ren ondorio orokorrak, emaitzak hobeak izan zitezkeela da. Baina, hala ere, hasiera batean, beste testuinguru batzuekin proba egitean, oso emaitza

txarrak eman zituen, hauek baina askoz okerragoak. Horrek argi uzten du, *UKB*-k testuinguru oso laburrarekin, zailtasun gehiago dituela.

Train eta Test-en banaketa

Beste ondorioetako bat, proiektuan zehar izandako oztopoen saihestea da. Hasiera batean, bai test eta train banatzean, intentzioa laburduraren agerpen bakoitzeko esaldika banatzea zen. Modu horretan banatu ziren bai test eta train, eta ondoren, esaldi bakoitzean hainbat laburdura egon daitezkeela konturatu ginen eta beraz, banaketa esaldi/paragrafoka egin zela esan behar da. Horrek esan nahi du, laburdura batzuetan corpusean agerpen gehiago daudela laburdurenak beste batzuetan baino. Azken datu hauek, estalduraren emaitzetan ikusi ahal izan dira.

Corpus-a banatzean gainera, etiketatzaileei agindu zitzaien bezala, puntuazio artean dihoazten laburdurak ez dira etiketatu behar eta, kasu horiek saihestu zirelako, laburdura batzuen kasuan, hasiera batean estimatu zena baino laburdura agerpen gutxiago zeuden. Adibidez, *C*-ren kasuan, A etiketatzaileari trainerako 300 laburdura zituen corpusa banatu zitzaion, baina etiketazio arau hori medio, azkenik 277 laburdura daudela esan behar da.

Ondorio gehiagoren artean, aurreko kasuan azaldu den etiketazio araua medio, etiketatzaileek kasuren batean ez dute kasu egin, eta etiketatu behar ez zen laburduraren bat etiketatu dute eta beste kasuren batean, laburduraren bat etiketatzea pasa zaie. Hori normala da eta giza-akatsa bezala onartzen da. Etiketatze lana, errepikorra delako, oso lan gogorra da eta X esaldi etiketatzean normala da nekea medio, laburdura batzuk galtzea.

7.1.2 Ondorio pertsonalak

Proiektu hau amaitzean, hainbat ondorio pertsonal atera ditut. Alde batetik, proiektua hasiera batean uste nuena baina askoz gehiago luzatu zait. Hori gertatzearen arrazoi handienetako bat, laugarren mailan, bigarren lauhilekoan, proiektua hastean, beste 4 irakasgai edukitzea izan da, eta honek nahigabe, proiektua beste urte betez luzatzea eragin du.

Pertsonalki, proiektuari dagokionean, hainbat gauza berri ikasi eta esperientzia jaso dudala esan daiteke. Hasiera batean ez nuen ideia handirik, ez corpus-ak ezta laburduren desanbiguazioari buruz.

Lanean aurrera egin ahala, hasiera batean pentsaezinak nituen gauzak ikasi ditut. Alde batetik, *UKB* sistema ikasi dut erabiltzen eta horrez gain, honek behar dituen fitxategiak prozesatzen eta osatzen ikasi dut. Prozesaketa horien artean, *KAF* fitxategien inguruan ikertu behar izan dut, eta horrez gain, ikertzen ikasi dut.

Laburduren desanbiguazioan ematen diren hainbat pausu egiten ere ikasi dut: Train eta Test zertarako erabiltzen diren, etiketatzaile batzuekin lana egiten, eta baita beste profesional batzuekin kontaktuan jartzen arazoei aurre egiten jakiteko.

Esan beharra dago, hasiera batean, erabiltzen nituen kontzeptuak, bai corpusa, maiztasunak edo hiztegia, ez nituela oso ondo ulertzen. Baina, proiektua egiten joan ahala, puzzleko piezak osatzen joan naizela esan daiteke, eta pieza bakoitzak zein funtzionalitate zuen ulertzen. Ikasketa prozesu jarrai bat izan dela esan daiteke.

Proiektuan ikasitakoaz gain, alde pertsonalean, esan beharra dago batzuetan, gogorra egin zaitela aurrera jarraitzea proiektuarekin, hainbat oztopo handi izan ditugulako. Pazientzia handia behar duen lana da proiektu hau, eta batzuetan, pazientzia hori galdu egiten da, baina ikasi egin behar da, gauzak lasai hartuz aurrera joaten. Alde honetik begiratuz, oztopoei aurre egiten ikasi dut eta hauei soluzio bat bilatzen.

Azkenerako, proiektua luzea egin dela esan beharra da, beti hobetzen joan daitekeen proiektu mota bat delako, eta beti egongo direlako zuzendu beharreko atazak, sistema on bat lortzeko.

7.2 Etorkizunerako lana

Proiektu hau amaitu arren, badira etorkizunari begira landu ahal daitezkeen hainbat lan eta atal. Jarraian, nire ustetan hobekuntzari dagokionean, ondo etorriko litzatezkeen hainbat aipamen egingo ditut. Horretarako, proiektuan zehar eman diren emaitzetan eta baita esperientzian oinarrituko naiz:

7.2.1 Corpus desberdinekin proba

Emaitzeri begirada bat emanik, oso argi geratu da, erabili den corpusean, hainbat laburdura ez direla anbiguo, *SEDOM*(dokumentazio klinikoko laburduren hiztegia)-n anbiguo izan arren. Horren adibide argi bat, *DM* laburdura

da, zein, *DM tipo II* agertzen den corpuseko agerpenetan eta *Freeling-med* berak desanbiguatu duen bere testuinguruaren arabera, *Diabetes melitus* dela esanez.

Baina, *DM* laburdurak gogoratu behar da, 4 hedapen posible dituela *SEDOM* hiztegiaren arabera:

- Densitometría
- Dermatomiositis
- Diabetes mellitus
- Duramadre

Hau ikusirik, beste corpus bat osatuz, probatu ahalko litzateke ea *DM* beste corpus horietan ere ez-anbiguo ala anbiguoa den.

Beste corpus berri horren osaketak, beste hainbat laburdura anbiguo ala ez-anbiguo diren jakiteko ere balioko luke, adibidez, *h* zein proiektuko corpusaren arabera, beti *'hora'* den. GAP honetan zehar erabilitako programa guzti-guztiak erabili ahalko lirateke (parametro batzuk aldatuta soilik) corpus berri batekin probak egiteko.

7.2.2 Laburdurak anbiguo desberdinen aukeraketa

Corpus desberdin bat osatu beharrean laburdura berak probatzen joateko, beste posibilitate bat, laburdura desberdinak corpus berberarekin probatzea izango litzateke, edo baita ere, beste corpus batekin. Hau da, 18 laburdura izan dira aukeratuak proiektuan lantzen joateko eta horietatik hainbat, soilik letra batekoak ziren (anbiguo ez direnak suertatu). Laburdura horiek eman dituzte orokorrean bai estalduran eta doitasunean emaitza txarrenak.

Hobekuntzak eman ahal izateko hainbat erregela jarri ahalko lirateke, probatuko diren laburdurak aukeratzeko. Horietako erregela bat, laburduren letra kopurua kontuan hartzea litzateke. Interesgarria izango litzateke, laburduren letra kopuruen arabera, ikustea zein diren bai estaldura eta baita ere doitasunen portzentaiak.

Hobe azalduta, ondo egongo litzateke, adibidez, *h* , *K* , *T* , *C* ,... bezalako laburdurak alde batetik, beste batetik *CD,DA,DM...* modukoak eta beste batetik *EAC* , *ECG* , *HTA...*-ren moduko laburdurak probatzen joatea, eta

emaitzak ikustea. Zein letra kopuruko laburdurek dituzten emaitza hoberenak jakin ahalko genuke.

Ziurrenik, letra kopuru bakarreko laburdurek ez-anbiguoen gisako emaitzak emango lituzteketeela ia corpus gehienetan, baina probatzea ez litzateke gaizki egongo ziur egon ahal izateko.

7.2.3 Beste teknika batzuen erabilera

Proiektuan zehar, teknika gisa, *UKB* , ausazkotasunean oinarritutako baslinea eta maiztasunean oinarritutako sistemak erabili dira. Baina, hizkuntzaren prozesamendua aurrera doan arlo bat da eta teknika berriak ari dira sortzen, prozesamendurako. Baliteke, hizkuntzaren prozesamenduan erabili ohi diren teknika ugari medikuntza arloan ere erabili ahal izatea.

7.2.4 Hiztegiaren hobekuntza

UKB erabili ahal izateko, parametro gisa, grafoa, hiztegia eta testuingurua pasa behar zaizkio. Hiztegian azaldu den bezala, bai hedapenak beraien CUI-ekin eta baita laburdurak hainbat CUI-ekin ere agertzen dira. Baina, kasu batzuetan, gertatu da, hiztegiak, ez ekartzea laburdura bakoitzak dituen hedapen posible guztien CUI-ak, eta horrek, erroreak sortu ahal izan ditu doitasunean.

Beraz, hobekuntza gisa, ondo egongo litzateke hiztegia osatzen joatea falta zaizkio hedapen eta beraien CUI-ekin. Modu honetan, ziur emaitzak *UKB*-rekin hobeak izango liratekela.

7.2.5 Esaldi luze eta txikien banaketa

Proiektuan zehar, corpuserako esaldien banaketa egin zen, txiki eta handien artean, hau da, 5 hitz baino txikiago eta 5 hitz baino gehiago zituzten esaldien artean. Banaketa hau, *UKB*-k testuinguru motz edo handi batekin nola funtzionatzen zuen ikusteko egin zen, baina azkenean ez zen probatu. Hau da, azkenean, esaldi guztiak, bai txikiak eta baita luzeak ere, batera probatu ziren.

Beraz, hobekuntza gisa, etorkizunerako lan posible baterako, ongi egongo litzateke, esaldi luzeak alde batetik, eta beste batetik txikiak, probatzea *UKB*-

rekin, ikusi ahal izateko, testuinguru zabal edo motz batek nolako eragina duen.

Kapitulua 8

Jarraipen eta Kontrola

Ataza honetan, proiektuan zehar egin den jarraipen eta kontrola azalduko da. Horretarako, denboren estimazioak azaltzen diren irudi bat eta datuen laburpen bat erakutsiko dira.

8.1 Denbora-estimazioak

Jarraian ataza eta azpiataza bakoitzari esleitutako denbora estimazioak aurkezten dira.

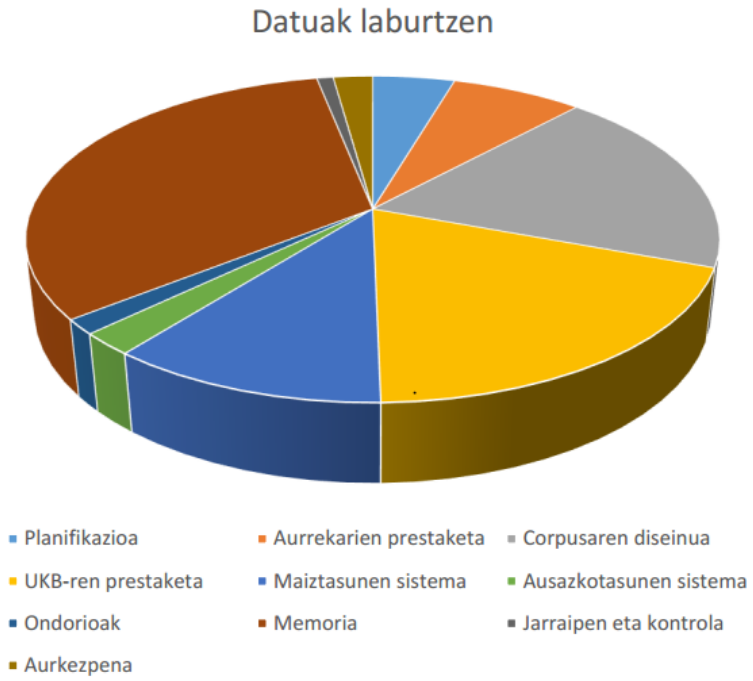
Denboren estimazioa irudian ikusten den moduan, hasiera batean planifikatu zen denboraren oso desberdina suertatu da prozesua. Bi ataletan, behar izan da denbora gehien: corpusaren diseinuan eta UKB erabili ahal izateko sortu behar izan diren fitxategi, programa, analisi... guztietan.

ATAZA	ESTIMATUTAKO DENBORA (ORDUAK)
Planifikazioa	25 ordu
Aurrekarien prestaketa eta analisia	40 ordu
Corpusaren Diseinua	103 ordu
Analisia	20 ordu
Diseinua	15 ordu
Inplementazioa	45 ordu
Probak	13 ordu
Ebaluazioa	10 ordu
UKB-ren prestaketa	104 ordu
Analisia	15 ordu
Hiztegiaren osaketa	26 ordu
Kontestuaren osaketa	20 ordu
Inplementazioa	23 ordu
Probak	12 ordu
Ebaluazioa	8 ordu
Maiztasunen sistema	60 ordu
Analisia	13 ordu
Inplementazioa	22 ordu
Probak	20 ordu
Ebaluazioa	5 ordu
Ausazkotasunen sistema	12 ordu
Inplementazio	12 ordu
Ondorioak	8 ordu
Memoria	62 ordu
Jarraipen eta Kontrola	5 ordu
Aurkezpena	12 ordu
Guztira	**431 ordu**

Irudia 8.1: Denboren estimazioa

8.2 Datuak laburtzen

Atal honetan, datuen laburpen bat erakusten duen irudi bat azalduko da. Irudi honetan, ikusgai egongo dira, zein ataletan, behar izan den esfortzu eta denbora gehiago.



Irudia 8.2: Datuak laburtzen

Denbora estimazioen irudiari erreparatzen badiogu, proiektuan denbora gehien behar izan diren atazak memoriaren diseinuaz gain, garapenera dedikaturiko moduluak direla ikusten da, corpusaren diseinua eta UKB-ren prestaketa hain zuen ere.

Atal guztietan, analisiak izan du lan karga handia, ikerketa aldeko proiektu bat denez, analisi eta ikerketa ugari egin behar direlako. Hau da, bai UKB erabiltzeko, teknika berriak ikasteko...

Kapitulua 9

Bibliografia

1. Euskarazko wikipedia. URL: http://eu.wikipedia.org/.
2. Ixa ikerketa taldea. URL: http://ixa.si.ehu.es.
3. SEDOM hiztegia. URL: http://www.sedom.es/diccionario/.
4. SNOMED CT.

   URL: http://www.msssi.gob.es/profesionales/hcdsns/areaRecursosSem/snomed-ct/quees.htm.
5. UMLS. URL: https://www.nlm.nih.gov/research/umls/quickstart.html.
6. Disambiguation of Biomedical Abbreviations (Mark Stevenson1, Yikun Guo, Abdulaziz Al Amri and Robert Gaizauskas).

   URL: http://anthology.aclweb.org/W/W09/W09-13.pdf#page=83.
7. Automatic annotation of medical records in spanish with disease drug and substance names.

   URL: http://ixa.si.ehu.es.
8. Graph Based Word Sense Disambiguation and Similarity.

   URL: http://ixa2.si.ehu.es/ukb/.
9. UKB. URL: http://ixa2.si.ehu.es/ukb/.
10. BRAT. URL: http://brat.nlplab.org/.
11. PERL.

URL: http://www.unibertsitatea.net/blogak/testuak-lantzen/2009/08/31/22-kasu-praktikoa-n-karaktereko-hitzak/.

12. Word Sense Disambiguation.

    URL: https://en.wikipedia.org/wiki/Word-sense_disambiguation.

13. Word Sense Disambiguation: A Survey.

    URL: http://www.ling.upenn.edu/courses/ling052/Navigli2009.pdf.

14. Machine Learning Techniques for Word Sense Disambiguation.

    URL: http://www.cs.upc.edu/ escudero/wsd/06-tesi.pdf.

15. Knowledge-based biomedical word sense disambiguation: comparison of approaches.

    URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-569.

16. Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain.

    URL: http://www.d.umn.edu/ tpederse/Pubs/amia07.pdf

17. UKB.

    URL:https://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1233561923/publikoak/ukb_eacl09

18. Maite Oronoz, Arantza Casillas, Koldo Gojenola eta Alicia Perez 2013 Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013 Havana, Cuba, November 20-23, 2013 Proceedings, Part II

19. Bosma W. E., Vossen P., Soroa A., Rigau G., Tesconi M., Marchetti A., Monachini M. and Aliprandi C.. KAF: a generic semantic annotation format. In: Proceedings of the Generative Lexicon (GL2009) Workshop on Semantic Annotation, Pisa, Italy, 2009.