

Títol: Mòdul de correferència per al software freeling

Volum: 1/1

Alumne: José Miguel Collado Fructuoso

Director/Ponent: Horacio Rodríguez Hontoria

Departament: Llenguatges i Sistemes Informàtics

Data: 26-01-2009

DADES DEL PROJECTE

Títol del Projecte: Mòdul de correferència per al software freeling

Nom de l'estudiant: JOSE MIGUEL COLLADO FRUCTUOSO

Titulació: ENGINYERIA EN INFORMÀTICA

Crèdits: 37.5

Director/Ponent: HORACIO RODRÍGUEZ HONTORIA

Departament: Llenguatges i Sistemes Informàtics

MEMBRES DEL TRIBUNAL (nom i signatura)

President:

Vocal:

Secretari:

QUALIFICACIÓ

Qualificació numèrica:

Qualificació descriptiva:

Data:

Índex

1	Introducció	7
1.1	Que és la coreferència	7
1.2	Objectius	8
1.3	Aplicacions	9
2	Treball previ	11
2.1	Proposta de Wee Meng Soon, Hwee Tou Ng i Daniel Chung Young Lim	11
2.2	Adaptació al castellà	13
3	Eines utilitzades	15
3.1	Freeling	15
3.2	CESS i Ancora	16
3.3	Fries i Omlet	19
4	Arquitectura de la solució proposada	21
4.1	Visió general	21
4.2	Aprenentatge	22
4.3	Descripció del Mòdul	22
4.3.1	Parser del corpus <i>Ancora</i> i <i>CESS</i>	22
4.3.2	Codificador i Extracció de característiques	24
4.3.3	Entrenament dels models	26
4.3.4	Validació dels models	26

4.3.5	Integració a <i>Freeling</i>	27
5	Experiments i resultats	33
5.1	Condicions inicials.	33
5.1.1	Generació d'exemples positius i negatius	33
5.1.2	Extracció de característiques	34
5.1.3	Resultats obtinguts	34
5.2	Limitació de exemples negatius i expansió dels positius	34
5.3	Problemes amb els exemples originals. Solucions	35
5.4	Característiques més rellevants	38
5.5	Corbes d'aprenentatge	42
6	Planificació i costos	49
6.1	Planificació inicial	49
6.2	Imprevistos i planificació real	50
6.3	Costos	51
7	Conclusions	53
7.1	Problemes i possibles solucions	53
7.1.1	Problema 1. Positius vs Negatius	53
7.1.2	Solució 1. Enriquir el corpus <i>AnCora</i>	54
7.1.3	Problema 2. Característiques que no aporten res	56
7.1.4	Problema 3. Extracció de <i>SN</i> al <i>Freeling</i>	57
7.1.5	Millora 1. Característiques	57
7.1.6	Millora 2. Entrenament	57
A	Conflictes a <i>AnCora</i>	59
B	Glossari	65
	Bibliografia	69

Índex de taules

5.1	Taula amb les quantitats de negatius i positius.	35
5.2	Taula de resultats.	36
5.3	Taula amb les quantitats de negatius i positius amb nous negatius.	37
5.4	Taula de resultats amb negatius ampliats sobre els models anteriors.	38
5.5	Taula de resultats amb negatius ampliats sobre els propis models.	39
6.1	Planificació inicial.	49
6.2	Planificació final.	51
6.3	Perfils necessaris.	51
6.4	Cost del projecte.	51

Índex de figures

1.1	Relacions de <i>referència</i> i <i>coreferència</i> entre el món real i el món textual.	7
2.1	Arquitectura del processament del llenguatge natural	11
3.1	Demo online de la web de <i>Freeling</i>	16
4.1	Diagrama de processos. Mòduls desenvolupats	21
4.2	Tipus de dades al <i>Freeling</i>	28
4.3	Diagrama de processos al <i>Freeling</i> amb el nou mòdul <i>coref</i>	29
5.1	Variacions entrenant el model eliminant una a una les característiques.	40
5.2	Variacions entrenant el model sense el <i>número</i> eliminant una a una les característiques.	40
5.3	Corbes d'aprenentatge sobre F-score amb totes les característiques.	42
5.4	Corbes d'aprenentatge sobre F-score únicament amb la característica de distància.	43
5.5	Corbes d'aprenentatge sobre Precision amb totes les característiques.	44
5.6	Corbes d'aprenentatge sobre Precision únicament amb la característica de distància.	45
5.7	Corbes d'aprenentatge sobre Recall amb totes les característiques.	46
5.8	Corbes d'aprenentatge sobre Recall únicament amb la característica de distància.	47

Capítol 1

Introducció

1.1 Que és la coreferència

Quan una unitat lingüística ens serveix per referir-nos a una entitat diem que hi ha una relació de referència entre l'unitat lingüística i l'entitat. Quan dues unitats lingüístiques es refereixen a la mateixa entitat diem que entre elles hi ha una relació de coreferència (veure la Figura 1.1).

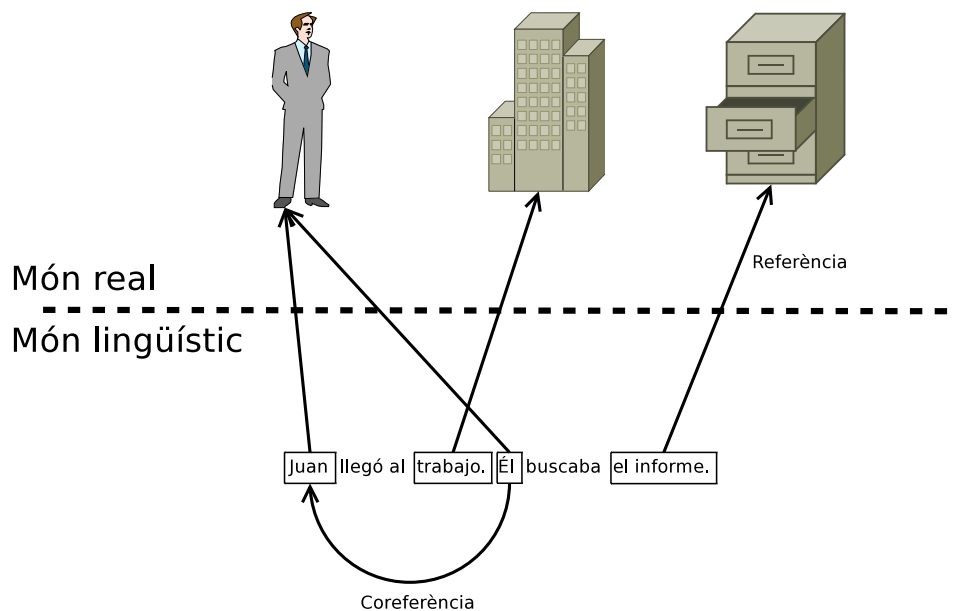


Figura 1.1: Relacions de *referència* i *coreferència* entre el món real i el món textual.

Les diferents unitats lingüístiques que en un text es refereixen a entitats del món real s'anomenen mencions (mentions). Les mencions són sempre sintagmes nominals. Poden ser:

Nominals: Un sintagma nominal governat per un nom comú, normalment el sintagma nominal és definit, *el presidente del gobierno*.

Entitats nombrades: (Named entities) *José Luis Rodríguez Zapatero*.

Pronominals: Pronom personal, demostratiu, possessiu, etcètera.

Hi han dos tipus de coreferència principals:

Coreferència indirecta: “he visto una nueva casa ayer. La cocina era excepcionalmente grande” (la cocina = de la casa).

Coreferència directa: “he visto una nueva casa ayer. Su cocina era excepcionalmente grande” (su = de la casa), exemples de (Chafe 1974).

En el cas de la coreferència directa, es pot trobar dos casos diferents segons el lloc on apareix la primera menció a la entitat del món real:

Anafòric: Quan el sintagma nominal que fa la primera menció a la entitat (nominal o entitat nombrada) apareix abans en el text que el coreferent.

Catàfòric: Quan el sintagma nominal apareix després del coreferent. La catàfora normalment és un recurs literari i poc freqüent.

La resolució automàtica de la coreferència es tracta de marcar en un document les relacions de coreferència que apareixen de forma implícita. Aquesta és una de les tasques més difícils que hi han dins del processament del llenguatge natural. Bàsicament hi han dos estratègies. Una és creant una sèrie de regles partint de diferents informacions com morfològica, sintàctica, semàntica, anàlisi del discurs, un diccionari d'escenaris, etcètera. Normalment aquesta tasca és duta a terme per lingüistes. L'altra és fent aprenentatge automàtic a partir de la extracció de informació dels constituents, partint d'un corpus etiquetat. En aquest projecte s'ha optat per aquesta última estratègia.

1.2 Objectius

L'objectiu és crear un mòdul de coreferència per al software *Freeling* <http://www.lsi.upc.edu/~nlp/freeling> ([8] i [9]) amb màxim percentatge d'èxit possible.

Freeling és un software *open source* que conté un conjunt d'eines lingüístiques com ara anàlisi morfològic, anàlisi sintàctic, classificació d'entitats, nombrades, etcètera. ¹

La idea és que s'integri plenament en els tipus de dades ja definits al *Freeling*, afegint els tipus de dades necessaris. En aquest cas la estratègia escollida és fer aprenentatge automàtic a partir d'un corpus etiquetat manualment i crear un conjunt d'arbres de decisió que després el *Freeling* farà servir per respondre si dos elements sintàctics són coreferents o no. Per aquesta solució, utilitzarem la informació d'anàlisi morfològic que ens dona *Freeling*, un anàlisi sintàctic superficial i informació semàntica procedent de *Wordnet* també integrat al *Freeling*. A partir d'aquesta informació es generaran uns conjunts d'exemples amb els que es construirà un model d'aprenentatge que s'integrarà al *Freeling*. Després es programarà al *Freeling* el mòdul corresponent per fer les consultes al model i respondre si dos sintagmes són coreferents o no.

Partim d'una proposta per la llengua anglesa anomenada *A Machine Learning Approach to Coreference Resolution of Noun Phrases*, creada per *Wee Meng Soon, Hwee Tou Ng i Daniel Chung Young Lim*. Es tracta d'una aproximació senzilla però que dona resultats excel·lents i ha estat sovint utilitzada com a *baseline* per molts sistemes més sofisticats de resolució de la coreferència. L'objectiu d'aquest treball és adaptar aquesta proposta al castellà aconseguint uns resultats similars i integrat al *Freeling* com un mòdul més. Val a dir, que la mateix desenvolupament es podrà utilitzar en català i d'altres idiomes que estiguin suportats al *Freeling*. També s'ha tingut en compta la tesis de la Marta Recasens *Towards Coreference Resolution for Catalan and Spanish*. [10]. En aquesta tesis hi ha una àmplia bibliografia sobre coreferència.

1.3 Aplicacions

La resolució de la coreferència és una tasca molt important del processament del llenguatge natural (NLP). És necessària en una àmplia gama de tasques de NLP, de la comprensió del llenguatge a la estadística, l'anàlisi del discurs, la extracció d'informació, la traducció i la elaboració de resums.

¹veure secció 3.1

Capítol 2

Treball previ

2.1 Proposta de Wee Meng Soon, Hwee Tou Ng i Daniel Chung Young Lim

La proposta *A Machine Learning Approach to Coreference Resolution of Noun Phrases*, ofereix un mètode d'aprenentatge automàtic a partir d'un corpus etiquetat en anglès. El mètode proposat consisteix primer en processar el text per obtenir finalment els candidats a coreferents.

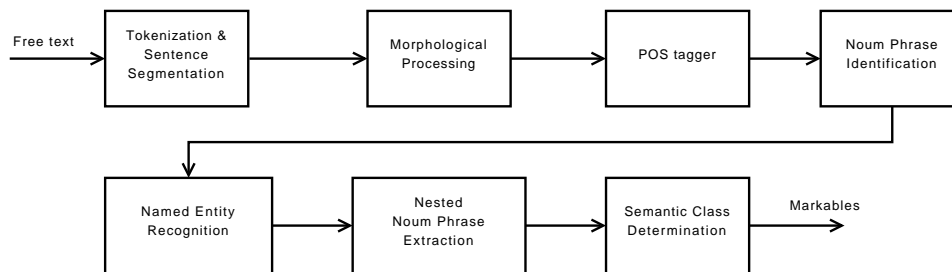


Figura 2.1: Arquitectura del processament del llenguatge natural

Una vegada seleccionat el conjunt de candidats, s'extreu per cada parella un vector de característiques codificades numèricament. En concret proposen 12 característiques per cada parella i i j , on i és el possible antecedent i j és la possible anàfora:

1. **Distance.** Distància en sentències. 0 si estan en la mateixa sentència, 1 si estan a una sentència de distància, etcètera.
2. **i-Pronoun.** Cert si i és un pronom.
3. **j-Pronoun.** Cert si j és un pronom.

4. **String Match.** Cert si i i j són iguals després d'eliminar els determinants i els pronoms demostratius.
5. **Definite Noun Phrase.** Cert si j comença per un article definit (el, la, els...)
6. **Demonstrative Noun Phrase.** Cert si j comença per un adjectiu demostratiu.
7. **Number Agreement.** Cert si i i j coincideixen en número (singular o plural).
8. **Semantic Class Agreement.** Cert si i i j pertanyen a la mateixa classe semàntica en un conjunt fixat (persona, organització, diners, data, ...).
9. **Gender Agreement.** Cert si i i i coincideixen en gènere (masculí o femení).
10. **Both Proper Names.** Cert si tant i com j són noms propis.
11. **Alias.** Cert si j es pot considerar un *alias* de i . Es considera que poden ser *alias* si j és un acrònim de i , o si i i j coincideixen en una part del prefixe.
12. **Appositive.** Cert si j és una aposició de i .

Per fer l'aprenentatge, es consideren les cadenes de coreferents dins del corpus etiquetat i es generen exemples positius (que corefereixen) i negatius. Els negatius, es generen només dins de les pròpies cadenes. Per exemple, en el cas $A1 - b - c - A2 - d - e - A3$ on les A^* són una cadena de coreferents i la resta són sintagmes nominals que no corefereixen en aquesta cadena, es generen com exemples positius $(A1, A2)$ i $(A2, A3)$, i com exemples negatius $(b, A2)$, $(c, A2)$, $(d, A3)$ i $(e, A3)$.

Una vegada seleccionat el conjunt d'exemples, es genera un arbre de decisió amb l'algorisme C5, que és el que després es consultara per respondre a la pregunta de si dos sintagmes nominals corefereixen o no.

Una vegada construït el sistema i utilitzant els corpus MUC-6 i MUC-7 s'aconsegueix els següents resultats:

MUC-6: Recall de 58.6%, Precisió de 67.3% i un F-score de 62.6%

MUC-7: Recall de 56.1%, Precisió de 65.5% i un F-score de 60.4%

2.2 Adaptació al castellà

El propòsit d'aquest projecte és adaptar el treball de *Soon et al.* ([6]) al castellà. Aquesta adaptació es podria utilitzar també per al català (es deixa com un treball futur) donat que fem servir un corpus etiquetat que ja existeix en les dues llengües (*l'AnCora*). Partim de les mateixes característiques, però per tal d'integrar-lo com a mòdul dins del software *Freeling* canviem l'algorisme d'aprenentatge *C5* per *l'Adaboost* (*Shaphire & Freund [7]*), que és el que fa servir el *Freeling* per defecte. També fem servir per extreure la informació semàntica *EuroWordnet*, que també ens la proporciona *Freeling*.

Capítol 3

Eines utilitzades

3.1 Freeling

El software *Freeling* (<http://www.lsi.upc.edu/~nlp/freeling>) [9] és una llibreria que proveeix un conjunt d'eines d'anàlisi de llenguatge natural. Està dissenyat per funcionar com una llibreria externa per a qualsevol aplicació que requereixi l'ús de tractament del llenguatge. També ofereix un simple programa per línia de comandes per poder interaccionar amb la llibreria i poder fer ràpidament l'anàlisi de fitxers de text.

Les seves principals funcionalitats són:

- Separació de paraules
- Separació de sentències
- Anàlisi morfològic
- Tractament de sufixos
- Reconeixement de multiparaules
- Separació de contraccions
- Predicció probabilística de paraules desconegudes
- Detecció d'entitats nombrades
- Reconeixement de dates, números, ràtios, monedes i magnituds físiques (velocitat, massa, temperatura, densitat, etcètera.)
- PoS tagging
- Anàlisi sintàctic basat en charts

- Classificació d'entitats nombrades
- Marcat semàntic basat en wordnet
- Anàlisi de dependències basat en regles.

Aquestes funcionalitats estan disponibles per a tots els idiomes suportats: Castellà, Català, Gallec, Italià i Anglès. A la seva web hi ha una demostració online (<http://garraf.epsevg.upc.es/freeling/demo.php>) on es poden veure algunes de les seves funcionalitats. A la figura 3.1 podem veure un exemple.

FreeLing 2.1
AN OPEN-SOURCE SUITE OF LANGUAGE ANALYZERS

Write your sentences

El gato come pescado y bebe agua.
Yo bajo con el hombre bajo a tocar el bajo bajo la escalera.
Mi amigo Juan Mesa de mesa la barba al lado de la mesa.

Analysis options

- Multword detection
- Number recognition
- Date/Time recognition
- Quantities, ratios, and percentages
- Named Entity detection
- Named Entity classification
- No sense annotation
- WN sense annotation: All senses
- WN sense annotation: Most frequent sense

Select language: Spanish | Select output: PoS Tagging | Submit

Analysis Results

Sentence #1

El gato come pescado y bebe agua .
el gato comer pescado y beber agua .
DA0MS0 NCMS000 VMIP3S0 NCMS000 CC VMIP3S0 NCFS000 Fp

Sentence #2

Yo bajo con el hombre bajo a tocar el bajo bajo la escalera .
yo bajar con el hombre bajo a tocar el bajo bajo la escalera .
PP1CSN00 VMIP1S0 SPS00 DA0MS0 NCMS000 AQ0MS0 SPS00 VMN0000 DA0MS0 NCMS000 SPS00 DA0FS0 NCFS000 Fp

Sentence #3

Mi amigo Juan Mesa se mesa la barba a el lado de la mesa .
mi amigo juan_mesa se mesar el barba a el lado de la mesa .
DP1CSS NCMS000 NP000000 P0000000 VMIP3S0 DA0FS0 NCFS000 SPS00 DA0MS0 NCMS000 SPS00 DA0FS0 NCFS000 Fp

Figura 3.1: Demo online de la web de *Freeling*

3.2 CESS i Ancora

CESS [11] és un corpus d'arbres sintàctics de 500.000 paraules, anotat morfològicament i sintàcticament (constituents i funcions). També està disponible en format de dependències.¹

¹El corpus es pot descarregar des de <http://clic.ub.edu/ancora/?page=downloads.php>. Y es pot consultar via web des de: <http://clic.ub.edu/ancora?page=cerques.php>

AnCora [12]² és un corpus d'arbres sintàctics del català i de l'espanyol amb diferents nivells d'anotació:

- categoria morfològica
- constituents i funcions sintàctiques
- estructura argumental i papers temàtics
- classe semàntica verbal
- sentits de Wordnet nominals
- entitats nombrades

Com a resultat del procés d'anotació es disposa també de dos lèxics verbals de 2.580 entrades per a l'espanyol i 2.142 entrades per al català amb informació sobre la classe semàntica del verb i la subcategorització sintàctica, l'estructura argumental i els papers temàtics per a cada un dels seus sentits.

El corpus de cada llengua conté 500.000 paraules. Està constituït majoritàriament per textos periodístics. El corpus *AnCora* així com els lèxics verbals derivats *AnCora-Verb* estan disponibles.

Un exemple de marcat al corpus en format *XML*:

```
<sent file="palinka/es/CESS_ECE//1267_20000103.plnk.xml"
id="sent_1_9" src="1267_20000103_1.tbf">
  <w>Tras</w>
  <de id="de_114" type1="da0fs0" type2="nne" type3="SD">
    <w>la</w>
    <w>prórroga</w>
    <w>de</w>
    <de id="de_115" type1="da0fs0" type2="NE-other">
      <corefLink anchor="de_104" id="208" type="ident"/>
      <w>la</w>
      <w>Ley_de_Emergencia_Pesquera</w>
    </de>
    <de id="de_116" type1="rel" type2="spec">
      <corefLink anchor="de_114" id="221" type="ident"/>
      <w>que</w>
    </de>
    <w>se</w>
    <w>decretó</w>
  ...
```

²<http://clic.ub.edu/ancora/index.php>

Al corpus *AnCora* estan marcats tots els *DE*³. Podem veure com s'utilitza el *tag corefLink* per marcar els *DE*'s que són anàfora amb el seu referent. Per aquest projecte, el que ens interessa és extreure el marcat de *DE*'s per poder generar els exemples i de *corefLink* per tenir els exemples positius. Com es pot veure en aquest exemple, no hi ha la informació de *POS tagging*, i això és un greu problema, ja que es necessita aquesta informació. Per tal de recuperar aquesta informació primer es va intentar extreure el text per després passar-li al *Freeling* per tal de generar-la. Però no es va tenir èxit, ja que el problema de sincronitzar l'*Ancora* amb la sortida del *Freeling* es tornava massa complicada. La següent alternativa va ser extreure la informació del corpus CESS que conté els mateixos textos. Aquesta vegada sí que es va aconseguir marcar totes les paraules amb la seva informació i així tenir tot el que ens calia

³*Discourse Entities*, o entitats del discurs.

El mateix exemple corresponent a CESS:

```
(S
  (sp-CCT
    (prep
      (sps00 Tras tras))
    (sn
      (espec.fs
        (da0fs0 la el))
      (grup.nom.fs
        (ncfs000 pr rroga pr rroga)
      (sp
        (prep
          (sps00 de de))
        (sna
          (espec.fs
            (da0fs0 la el))
          (grup.nom.fs
            (np0000a Ley_de_Emergencia_Pesquera
              Ley_de_Emergencia_Pesquera))))
    (S.F.R
      (relatiu-SUJ
        (pr0cn000 que que))
      (morfema.verbal-PASS
        (p0000000 se se))
      (grup.verb
        (vmis3s0 decret  decretar))
    ...
```

Aqu  veiem com est  estructurat un fitxer *CESS*. La informaci  que s'extreu  s  nicament la categoria gramatical, ja que  s la mateixa que despr s generarem amb el *Freeling*.

3.3 Fries i Omlet

Fries  s una llibreria *open source* que ens permet convertir texts en llenguatge natural en vectors de caracter stiques pensats per servir com a *input* d'un algorisme d'aprenentatge autom tic. A m s, ens proporciona els tipus de dades per representar els documents. Veure la figura 4.2

Omlet  s una llibreria *open source* que ens ofereix eines orientades a l'aprenentatge autom tic. Cont  un extens *framework* on es poden implementar

nous algorithmes d'aprenentatge. En concret ens aporta l'algorisme *Adaboost* (*Shaphire & Freund [7]*), que és el que utilitzarem.

Capítol 4

Arquitectura de la solució proposada

4.1 Visió general

Aquesta solució és pot dividir en dues parts. El mòdul dins del software *Freeling*, que té la responsabilitat d'escollir dos sintagmes nominals i preguntar al mòdul *Adaboost* (*Shaphire & Freund [7]*) si són coreferents o no, i el mòdul que entrena el model que farà servir l'*Adaboost* (*Shaphire & Freund [7]*) a partir dels exemples generats del corpus *Ancora*.

A la figura 4.1 es veu una visió general dels mòduls desenvolupats en aquest projecte.

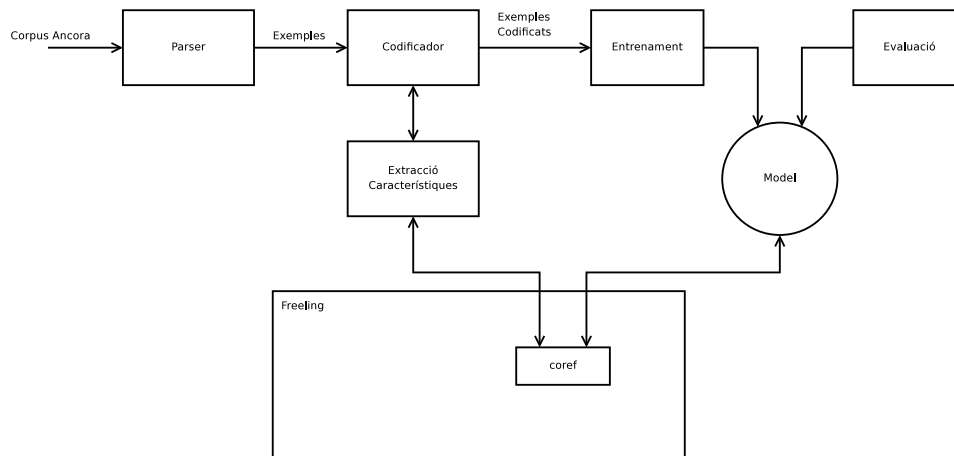


Figura 4.1: Diagrama de processos. Mòduls desenvolupats

4.2 Aprenentatge

En aquest projecte s'ha utilitzat una estratègia basada en *arbres de decisió*. Un *arbre de decisió* és un tipus d'arbre on cada node s'associa a una pregunta sobre una o varies característiques. Així amb un exemple i les seves característiques, l'arbre anant baixant sobre els seus nodes pot acabar responnent a la qüestió que ha après a respondre. Les fulles dels arbres acostumen a ser classes. Els *arbres de decisió* acostumen a basar-se en el principi d'entropia de la teoria de la informació.

En aquest cas s'ha utilitzat l'algorisme *Adaboost* (*Shaphire & Freund [7]*) (que ens el proporciona la llibreria *Omlet*). L'*Adaboost*, abreviació de *Adaptive Boosting*, és un algorisme d'aprenentatge creat per Yoav Freund i Robert Schapire. És un metaalgorisme i s'utilitza juntament amb altres algorismes classificadors més dèbils per formar un classificador més fort, el qual és el que dona la resposta final. És sensible al soroll, però també és més robust davant el sobreaprenentatge. A més és un algorisme adaptatiu, ja que cada classificador es crea en funció dels exemples no classificats prèviament.

El *Boosting* en que es basa *Adaboost* (*Shaphire & Freund [7]*) és un algorisme iteratiu on a cada iteració construeix un arbre aprofitant l'iteració anterior. Cada nou arbre intenta variant els pesos del conjunt d'exemples per beneficiar als casos mal classificats a l'iteració anterior.

Una cop tenim el conjunt d'arbres en el nostre model, la classificació d'un exemple nou es fa agafant les respostes que dona cada arbre i quedant-nos amb la més *forta*.

4.3 Descripció del Mòdul

4.3.1 Parser del corpus *Ancora* i *CESS*

L'objectiu d'aquest mòdul és generar un conjunt d'exemples positius i negatius a partir del corpus *Ancora*. S'ha de tenir en compte que aquests exemples no poden contenir cap informació que no doni el *Freeling*, ja que si féssim servir més informació, no es podria integrar després i esperar bons resultats. També de cara a poder fer diferents conjunts d'exemples, es va proveir a aquest mòdul de diferents paràmetres configurables. Una de les mancances que es va trobar al corpus *Ancora* és que no proveïa tota la informació necessària (com el Pos tagging complert, per exemple), és per això que al fer el parser, també es fa servir el recurs *CESS* de forma sincronitzada (són els mateixos articles).

La primera tasca a realitzar era com convertir el corpus *AnCora* del format

d'arbre en que està en una llista d'unitats lingüístiques. Hi havien diferents opcions; quedar-se només amb les més externes, o només amb les internes. Finalment es va optar per fer conservar tot, i cada *de* es conservava només el *head*.

Per generar el conjunt d'exemples positius, s'han implementat 2 algorismes. Considerem el següent cas on *A1*, *A2* i *A3* formen una cadena de coreferència, i *b*, *c*, *d* i *e* són sintagmes que no pertanyen a aquesta cadena (tot i que poden pertànyer a d'altres):

Per programar el procés dels fitxers XML, s'ha utilitzat la llibreria *libxml2* <http://xmlsoft.org/>

A1 b c A2 d e A3

El primer algorisme genera els exemples positius tal com es proposa a *Soon et al.* ([6]), és a dir, per parelles on cada anàfora es lliga amb el referent més pròxim. Així tindriem dos exemples positius: *A1-A2* i *A2-A3*. L'altre algorisme, el que fa és generar totes les combinacions possibles per cada cadena de coreferents, així afegiria com a tercer exemple *A1-A3*. El que es vol avaluar d'aquesta manera és si al tenir més exemples positius es poden obtenir millors resultats.

Per generar el conjunt d'exemples negatius, a *Soon et al.* ([6]) es proposa (sobre el mateix exemple) generar per cada parella positiva tants negatius com sintagmes hagin entre aquesta parella, i lligats a l'anàfora de la parella. És a dir, generaria: *b-A2*, *c-A2*, *d-A3*, *e-A3*. Ja amb previsió de que es podrien generar molts més negatius que positius, es van implementar uns filtres per tal d'eliminar negatius per distància o aleatòriament.

Finalment també es va implementar altres mètodes per generar negatius com per exemple el negatiu *d-e* degut als resultats. Això es veurà amb més detall al següent capítol junt amb els resultats.

Tots aquest exemples es generen en fitxers de text, per després ser processats. D'aquesta manera es poden generar de cop diferents conjunts amb diferents paràmetres una vegada i després fer-los servir segons convingui. Els fitxers generats es codifiquen de la següent manera:

+

0

0

2

El grupo estatal

```

da0ms0 ncms000 aq0cs0
0
4
6
-Fpa- EDF -Fpt-
Fpa np00o00 Fpt

```

On el primer caràcter és un $+$ o un $-$ segons sigui un exemple positiu o negatiu. Després tenim els dos elements del exemple amb aquesta informació: número de sentència, posició de la primera paraula, posició de l'última paraula, text del exemple i *PoS tagging* del text.

Amb aquesta informació extraurem les característiques de cada exemple.

4.3.2 Codificador i Extracció de característiques

Aquest mòdul té com a missió extreure les característiques de cada exemple que es faran servir tant per entrenar el model, com per consultar el model. La part que extreu les característiques pròpiament està en una llibreria i es crida des de la comanda que codifica els exemples de l'*AnCora* com des de el Freeling.

El codificador agafa per cada conjunt d'exemples generat amb el parser i genera en un fitxer els mateixos exemples codificats amb les seves característiques. Per la mateixa raó d'abans, codificar els exemples en fitxers intermedis, ens permet codificar una vegada i després fer diferents proves.

Les característiques extretes, són les mateixes que les que es proposa a *Soon et al.* ([6]) i d'altres noves on i és el possible antecedent i j és la possible anàfora:

1. **Sentència.** Distància en sentències. 0 si estan en la mateixa sentència, 1 si estan a una sentència de distància, etcètera.
2. **Distància.** Distància en paraules entre el final i i el principi de j . De 0 a n. **NOVA**
3. **i-Pronoun.** 1 si i és un pronom. 0 altrament.
4. **j-Pronoun.** 1 si j és un pronom. 0 altrament.
5. **i-Pronoun-tipus.** 1 si i és un pronom del tipus indicat. En total són 7 paràmetres segons el tipus de pronom (personal, demostratiu, possessiu, indefinit, interrogatiu, relatiu, exclamatiu). 0 altrament. **NOUS**

6. **j-Pronoun-tipus.** 1 si j és un pronom del tipus indicat. En total són 7 paràmetres segons el tipus de pronom (personal, demostratiu, possessiu, indefinit, interrogatiu, relatiu, exclamatiu). 0 altrament. **NOUS**
7. **String Match.** 1 si i i j són iguals després d'eliminar els signes, els determinants i els pronoms demostratius. 0 altrament.
8. **Definite Noun Phrase.** 1 si j comença per un article definit (el, la, els...). 0 altrament.
9. **Demonstrative Noun Phrase.** 1 si j comença per un adjectiu demostratiu. 0 altrament.
10. **Number Agreement.** 1 si i i j coincideixen en número (singular o plural). 0 altrament.
11. **Gender Agreement.** 1 si i i i coincideixen en gènere (masculí o femení). 0 altrament.
12. **Semantic Class Agreement.** 1 si i i j pertanyen a la mateixa classe semàntica en un conjunt fixat (persona, organització, diners, data, ...). 0 altrament. En aquest cas, s'ha diferenciat la manera d'aconseguir aquesta informació semàntica. En el cas dels noms comuns s'ha utilitzat les eines que proporciona el *Freeling* per accedir a la informació semàntica. En concret s'ha fet servir la ontologia de *EuroWordNet*. En el cas dels noms propis s'ha utilitzat la classificació de *Named Entities*. Evidentment quan s'ha de comparar un nom propi i un nom comú, es busca la equivalència entre les dues fonts.
13. **Both Proper Names.** 1 si tant i com j són noms propis. 0 altrament.
14. **Alias.** No usat en benefici dels següents vectors.
15. **Alias-acro.** 1 si j és pot considerar un acrònim de i . 0 altrament. **NOU**
16. **Alias-left.** 1 si j és un prefix de i . 0 altrament. **NOU**
17. **Alias-right.** 1 si j és un sufix de i . 0 altrament. **NOU**
18. **Alias-order.** 1 si i conté totes les paraules de j en el mateix ordre. 0 altrament. **NOUS**
19. **Appositive.** 1 si j és una aposició de i . 0 altrament.

Tots aquests resultats són parametrizables tant al codificador com al Free-ling, per així poder fàcilment configurar quins ens interessen més.

4.3.3 Entrenament dels models

L'entrenament dels models es fa amb l'algorisme *Adaboost* (*Shaphire & Freund [7]*), inclòs a la llibreria *Omlet*. Aquest algorisme genera un conjunt de arbres de decisió; cada un en base als exemples no classificats al anterior. Es pot parametritzar tant el número d'arbres màxim que volem i la profunditat dels arbres generats. A més ens dóna eines per veure a partir de quants arbres ja no aporten més informació. Això es pot veure clarament quan representem aquesta informació com una corba (*corba d'aprenentatge*). El model generat queda guardat en un fitxer que és el que en última instància farà servir *Freeling* per classificar els exemples nous.

4.3.4 Validació dels models

Per validar els models generats s'ha utilitzat la tècnica de la validació creuada o *Cross-Validation*. Consisteix en agafar els exemples per entrenar el model i dividir-los en grups (10 en aquest cas). Llavors s'agafen tots els grups menys un i s'entrena el model. Després és quantifica l'encert amb el grup que no s'ha fet servir. Això es fa tantes vegades com grups hem generat i després fem la mitja. Això ens permet tenir un resultat molt fiable, ja que si avaluem amb el mateix conjunt d'aprenentatge, ens arisquem a obtenir uns resultats més bons del que són en realitat.

Per fer la valoració de l'encert de cada model, utilitzem les mètriques de *precision*, *recall* i *F-score*, que són les que s'acostumen a fer servir en els sistemes basats en la recuperació de la informació.

precision. Ens diu de les respostes positives que dóna el model, quantes ho són realment. Un valor de 1.0 significa que totes les respostes positives són correctes. Es calcula com el número de positius marcats correctament dividit per el número de positius marcats

recall. Ens diu quantes respostes positives dóna el model respecte a les que haurien d'haver realment. Un valor de 1.0 significa que torna totes les respostes positives. Es calcula com el número de positius marcats correctament dividit per el número de positius esperats.

F-score. És la mitja harmònica ponderada entre la *precision* i el *recall*. Un valor de 1.0 significa que el model és capaç de reconèixer tots els exemples. En el cas particular on la ponderació dels dos valors té el mateix pes es calcula com:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.3.5 Integració a *Freeling*

Per fer l'integració al *Freeling*, primer s'han de crear els tipus de dades per tal d'emmagatzemar la informació de coreferència. Això s'ha fet tal com es veu a la figura 4.2 dins de la classe `document` declarada a la part de llenguatge de la llibreria *Fries*, en els fitxers `language.h` i `language.cc`.

La classe `document` queda així:

```
class document : public std::list<paragraph> {  
  
    paragraph title;  
    std::multimap<int, node> mapid;  
    std::map<node, int> mapnode;  
  
    int findGroup(node &node1);  
  
public:  
  
    document();  
    void add_positive(node &node1, node &node2);  
    int get_coref_id(const node &);  
    std::list<node> get_coref_nodes(const int id);  
    bool is_coref(const node &, const node &);  
};
```

On tenim:

- **mapid.** És una taula *hash* on tenim relacionats els *ids* de cada grup de coreferència amb tots els nodes (del *Freeling*) del document que pertanyen al grup.
- **mapnode.** És una taula *hash* on tenim relacionats els *nodes* del *Freeling* amb un identificador numèric que ens representa el grup de coreferència al que pertany.
- **add_positive.** És una funció constructora que insereix una parella que corefereix.
- **get_coref_id.** Funció consultora que ens retorna el *id* del grup de coreferència al que pertany el *node* o 0 si no pertany a cap.
- **get_coref_nodes.** Funció consultora que ens retorna tots els *nodes* que pertanyen al mateix grup de coreferència.

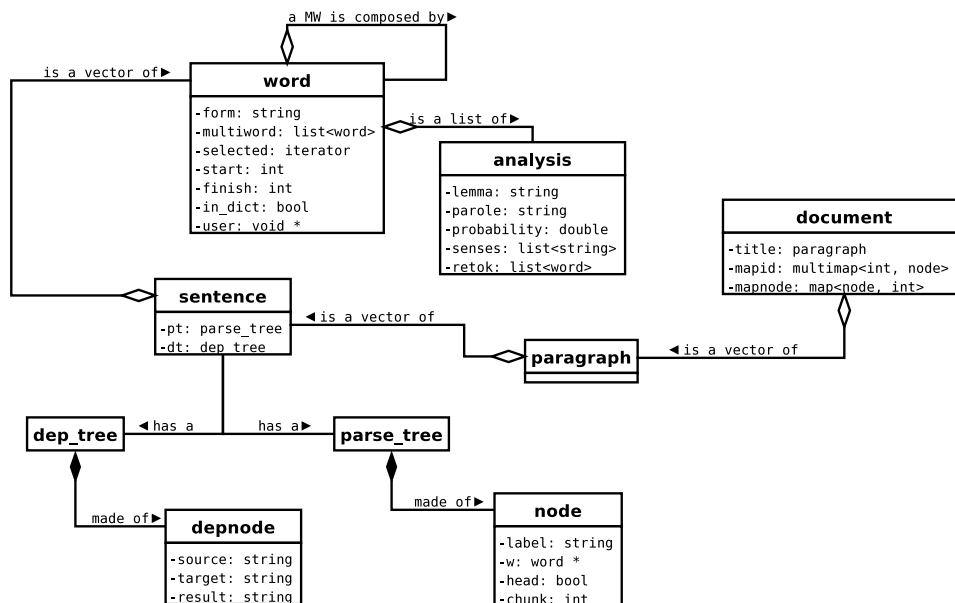


Figura 4.2: Tipus de dades al *Freeling*

- **is_coref.** Funció consultora que ens diu si dos *nodes* són coreferents o no.

Per un altra banda, tenim el component del *Freeling* pròpiament que té la responsabilitat de donat un document, escollir els candidats a coreferents, extreure les seves característiques, comprovar al model si són coreferents, i si és el cas, guardar-lo com a positiu. A la figura 4.3, es pot veure com quedaria el component junt amb la resta de components del *Freeling*. L'algorisme bàsicament és aquest:

Per cada sn en el document

Es busca el sn anterior com a candidat a referent.

Si és coreferent, es marca com a positiu

Sinó busquem un altre sn anterior fins una distància màxima

Es consideren tots els sintagmes nominals i per cada un es busca cap enrere si no és anàfora es mira el següent d'algun altre, fins una distància màxima també configurable.

Juntament amb aquests mòduls també s'ha desenvolupat un programa principal a mode d'exemple on se li passa un document, i fa tot el procés de crides i retorna els grups de coreferència.

Per resoldre la coreferència en un document s'utilitzen els següents mòduls:

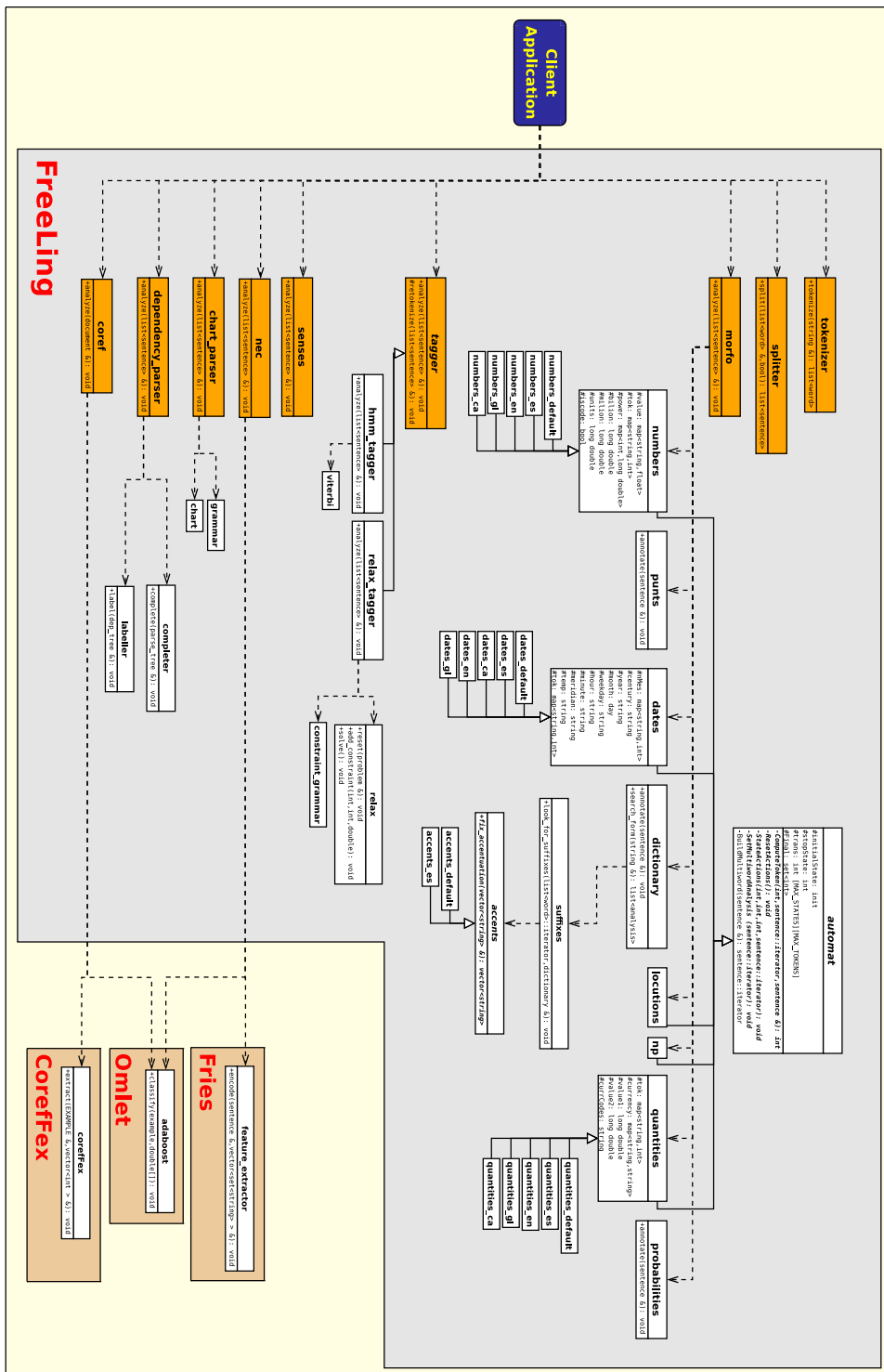


Figura 4.3: Diagrama de processos al Freeling amb el nou mòdul *coref*

- anàlisi morfològic
- *PoS Tagging*
- classificació de *Named Entities*
- anàlisi sintàctic
- cerca de coreferència

Tot hi que el *Freeling* és una llibreria d'eines i per tant no té una aplicació principal amb una interfície sofisticada, conté un parell de aplicacions simples que serveixen com exemples d'us de la llibreria. Per poder veure el resultat, s'ha creat una aplicació bàsica on fa les crides pertinents i mostra per consola el resultat. S'ha utilitzat el mateix tipus de bolcat que s'utilitza en els exemples afegint la etiqueta *REF #* a els *sn* que pertanyen a un grup de coreferents, on el número indica el grup al que pertanyen. Un fragment de sortida d'un document seria el següent:

```
S_[
  sn_[
    espec-ms_[
      +j-ms_[
        +(El el DAOMSO -)
      ]
    ]
    +grup-nom-ms_[
      +n-ms_[
        +(grupo grupo NCMS000 -)
      ]
      s-a-ms_[
        +a-ms_[
          +(estatal estatal AQOCSO -)
        ]
      ]
    ]
  ]
]

...

grup-sp_[
  +prep_[
    +(con con SPS00 -)
  ]
  sn(REF 10)_[
```

```

    espec-fs_[
      +indef-fs_[
        +(una uno DIOFS0 -)
      ]
    ]
  +grup-nom-fs_[
    +n-fs_[
      +(potencia potencia NCFS000 -)
    ]
  ]
]
sp-de_[
  +(de de SPS00 -)
  sn(REF 10)_[
    numero-nopart_[
      +(495 495 Z -)
    ]
    +grup-nom-mp_[
      +n-mp_[
        +(megavattios megavatio NCMP000 -)
      ]
    ]
  ]
]
]

```

On veiem com està marcat el grup de coreferents *10*. També podem veure que si un *sn* no s'ha detectat que sigui coreferent amb cap altre, llavors no porta cap marcatge especial.

Capítol 5

Experiments i resultats

En aquesta part es mostrarà com s'ha anat ajustant cada model per tal d'obtenir un resultat òptim. Val la pena destacar des de bon principi que les tècniques per eliminar negatius, encara que sobre el mateix conjunt d'exemples reduïts es comporten de forma molt acceptable (massa i tot), no ho fan tant bé quan s'enfronten en el conjunt original. Hi han decisions preses com a bones al principi del desenvolupament que al final s'han vist que no eren correctes i s'ha tingut de replantejar tot el procés. Per tal de fer més comprensible aquest procés, s'ha organitzat aquesta secció en funció de com han anat evolucionant les diferents parts amb els resultats obtinguts.

5.1 Condicions inicials.

5.1.1 Generació d'exemples positius i negatius

El corpus *Ancora* consta de 216 documents. En total hi han marcades 8593 relacions de coreferència. Seguint el mètode de *Soon et al.* ([6]) per generar tant els exemples positius com els negatius en tenim:

- 8593 exemples positius
- 158596 exemples negatius
- Un 5.13% de exemples positius sobre el total.

Aquesta desproporció entre els exemples negatius i els positius es deguda a que hi han exemples positius on el referent i l'anàfora estan molt lluny l'un de l'altre, llavors al generar els negatius genera tants com la distància en sintagmes. Per exemple:

$A1\ s1\ s2\ s3\ \dots\ sn\ A2$

Si considerem que A2 corefereix amb A1, llavors tindríem com a negatius $s1-A2$, $s2-A2$, $s3-A2$, i així fins $sn-A2$.

5.1.2 Extracció de característiques

Les característiques extretes són les assenyalades ja al capítol anterior. Es va optar per utilitzar totes les que venien donades per *Soon et al.* ([6]) i algunes afegides. La idea és treballar amb totes i al final intentar discriminar si hi han de més rellevants, irrelevants o inclús que afegeixen soroll (i per tant empitjoren el resultat final).

5.1.3 Resultats obtinguts

El problema és que quan s'entrena un model amb aquestes dades, fracassem, ja que genera un model que sempre respon *no* amb un error del 5% (els casos positius, és clar). Per tant, hem de corregir aquesta desproporció entre els exemples positius i negatius.

5.2 Limitació de exemples negatius i expansió dels positius

Per reduir aquesta diferència entre la quantitat d'exemples negatius i positius es poden utilitzar varies estratègies. Per poder comparar-les es van programar les següents opcions:

- La possibilitat de generar no només els exemples positius que proposa *Soon et al.* ([6]), sinó també totes les combinacions possibles de positius.
- Eliminar exemples negatius de forma aleatòria fins arribar a una proporció donada.
- Eliminar negatius a partir d'una distància llinar. La distància es mesura en sintagmes intermedis.

Fent diferents combinacions amb aquests paràmetres es van aconseguir els grups que es descriuen a la taula 5.1

En la taula 5.2 tenim el resultat d'entrenar els models i aplicar-los la validació creuada. Val la pena destacar el mal funcionament en els casos en

Filtres	Positius	Negatius	Positius/Negatius
Aplicant <i>Soon et al.</i> ([6])	8593	158596	5.13%
Negatius a distància màxima 20	8593	25685	25.06%
Negatius a distància màxima 10	8593	12881	40.01%
Negatius a distància màxima 5	8593	6089	58.52%
Eliminació de negatius aleatòria a proporció 1.0	8593	8592	50%
Tots els positius	26750	158596	14.43%
Tots els positius, Negatius a distància màxima 20	26750	25685	51.01%
Tots els positius, Negatius a distància màxima 10	26750	12881	67.49%
Tots els positius, Eliminació de negatius aleatòria a proporció 1.0	26750	26749	50%

Taula 5.1: Taula amb les quantitats de negatius i positius.

que no s'elimina cap negatiu. I comparant les dues formes d'eliminació de negatius, també es veu clarament que és més efectiu eliminar els negatius en funció de la distància, que no de forma aleatòria; per un altre banda té lògica, ja que els negatius més pròxims també són els més importants per no donar un fals positiu.

5.3 Problemes amb els exemples originals. Solucions

De tots els casos provats, es veu que en el cas de generar tots els positius i eliminat els negatius a una distància màxima de 10, és com s'obtenen els millors resultats. També és important veure que aquests resultats són sobre el propi conjunt d'exemples generats, que no necessàriament tenen perquè coincidir amb el conjunt de exemples que es mostrarien al model ja en un cas real (des de el *Freeling*). Si considerem sobre aquest cas:

$$A1 \ b \ c \ A2 \ d \ e \ A3$$

Podem pensar que per avaluar $A3$, com que tenim els exemples negatius $e, A3$ i $d, A3$ i el positiu $A2, A3$, l'algorisme d'aprenentatge té els exemples necessaris per poder aprendre a trobar $A2$. Però si considerem e , ens trobem

Filtres	Recall	Precision	F-Score
Aplicant <i>Soon et al.</i> ([6])	0.000000	0.000000	0.000000
Negatius a distància màxima 20	0.525659	0.978857	0.683464
Negatius a distància màxima 10	0.651938	0.999638	0.788699
Negatius a distància màxima 5	0.757463	0.972174	0.851335
Eliminació de negatius aleatòria a proporció 1.0	0.609691	0.571678	0.589389
Tots els positius	0.000888	0.335238	0.001770
Tots els positius, Negatius a distància màxima 20	0.758720	0.994178	0.860355
Tots els positius, Negatius a distància màxima 10	0.897061	0.968591	0.931418
Tots els positius, Eliminació de negatius aleatòria a proporció 1.0	0.636098	0.581784	0.607267

Taula 5.2: Taula de resultats.

que no hi ha cap exemple generat, ni positiu perquè no té, ni negatiu encara que de fet ho serien tots. Això feia sospitar que una vegada es busqués coreferència per *e*, o *d*... Possiblement donaria un fals positiu. Perquè l'algorisme implementat al *Freeling* (que és el proposat per *Soon et al.* ([6]) a la força ha de trobar molts més casos negatius que positius, ja que en un text normal és així.

Per veure millor com es comporta el model après davant de exemples més *realistes*, es generen nous grups d'exemples sota les següents condicions:

- Es generen els positius de les dues maneres provades.
- Es generen tots els negatius possibles, i limitats a una distància màxima.
- Es generen en el mateix ordre tots els conjunts, per tal de poder fer validació creuada amb els models anteriors.

A la taula 5.3 podem veure les quantitats de positius i negatius en cada conjunt de condicions escollides. El primer element és el resultat d'aplicar l'algorisme original de *Soon et al.* ([6]), per poder comparar amb els canvis fets. A la taula 5.4 tenim el resultat de validar els models anteriors amb el conjunt d'exemples ampliat amb una distància màxima de 5 (sempre agafant la part corresponent a la validació creuada, per no entrar en problemes de validar amb exemples ja vistos). A la taula 5.5 tenim el resultat de entrenar

Filtres	Positius	Negatius	Positius/Negatius
Aplicant <i>Soon et al. ([6])</i>	8593	158596	5.13%
Nous negatius a distància màxima 10	8593	278693	2.99%
Nous negatius a distància màxima 5	8593	141012	5.74%
Nous negatius a distància màxima 3	8593	84612	9.21%
Nous negatius a distància màxima 2	8593	56158	13.27%
Nous negatius a distància màxima 1	8593	27811	23.60%
Tots els positius. Nous negatius a distància màxima 10	26750	277683	8.78%
Tots els positius. Nous negatius a distància màxima 5	26750	140711	15.97%
Tots els positius. Nous negatius a distància màxima 3	26750	84488	24.04%
Tots els positius. Nous negatius a distància màxima 2	26750	56108	32.28%
Tots els positius. Nous negatius a distància màxima 1	26750	27811	49.02%

Taula 5.3: Taula amb les quantitats de negatius i positius amb nous negatius.

i validar els nous models. S'han utilitzat diferents limitacions per distància i tots els grups s'han validat amb els conjunts de distància 5 i 10.

Sembla que aquests resultats s'aproximen més al que ha de ser el resultat final de l'aplicació perquè es tenen en compte exemples que s'hauran de consultar i que abans no s'hi contava, com queda patent a la taula 5.4. Cal destacar que així ens apartem de la proposta de *Soon et al. ([6])*. Sobre els resultats generats, destacarem el cas on utilitzem tots els positius i els negatius limitats a distància 10. Una vegada validat, tenim un *F-score* de 0.414162. Es pot veure que si limitem la distància de validació podem obtenir millors resultats, però s'ha pres la decisió de utilitzar la millor combinació amb distància 10 per tenir un conjunt més real. També és important veure que obtenim una bona precisió en el resultat (0.89989 de precisió), però cobrim una part petita dels exemples (0.269577 de recall).

Filtres	Recall	Precision	F-Score
Negatius a distància màxima 20	0.525566	0.250415	0.338073
Negatius a distància màxima 10	0.651864	0.080288	0.142914
Negatius a distància màxima 5	0.757237	0.0582073	0.108082
Eliminació de negatius aleatòria a proporció 1.0	0.577857	0.0692656	0.123558
Tots els positius	0.000487448	0.3	0.000973145
Tots els positius, Negatius a distància màxima 20	0.611495	0.263238	0.367422
Tots els positius, Negatius a distància màxima 10	0.72273	0.0733927	0.133224
Tots els positius, Eliminació de negatius aleatòria a proporció 1.0	0.560714	0.0530734	0.096941

Taula 5.4: Taula de resultats amb negatius ampliat sobre els models anteriors.

5.4 Característiques més rellevants

Per tal de mesurar la rellevància que té cada característica en l'algorisme d'aprenentatge, s'ha optat per agafar el model més prometedor i tornar a repetir el procés d'aprenentatge eliminant les característiques d'una en una. Això ens permet detectar quines característiques són importants, quines no aporten res i fins hi tot si alguna aporta soroll i empitjora el resultat. També mostrem quina quantitat de casos tenim per cada característica. A la figura 5.1 podem veure com varia el *F-score*. Una de les coses que més crida l'atenció és que eliminant la informació sobre distància, el model és incapaç d'aprendre res, i dona un *F-score* de 0. En canvi la resta de característiques la variació és mínima. Això també sorprèn, perquè seria d'esperar que la variació fos més gran. Com a referència tenim el *F-score* amb totes les característiques.

En la figura 5.2 s'ha repetit l'experiment eliminant en tots els models la característica de *número*, ja que hem vist que era la que més millorava el resultat al ser eliminada. Tot hi així veiem que la variació continua sent mínima. Tenim com a referència el *F-score* del model sense el *número*. Ara ja veiem que no hi ha manera de millorar el resultat.

Filtres	Recall	Precision	F-Score
Negatius ampliat a distància 1. Validats amb distància 5	0.660943	0.0723766	0.130383
Negatius ampliat a distància 2. Validats amb distància 5	0.516683	0.136161	0.215373
Negatius ampliat a distància 3. Validats amb distància 5	0.429499	0.392985	0.408173
Negatius ampliat a distància 5. Validats amb distància 5	0.334159	0.940503	0.492965
Negatius ampliat a distància 1. Validats amb distància 10	0.660943	0.0263438	0.0506606
Negatius ampliat a distància 2. Validats amb distància 10	0.516683	0.0273656	0.0519709
Negatius ampliat a distància 3. Validats amb distància 10	0.429499	0.0333814	0.0619229
Negatius ampliat a distància 5. Validats amb distància 10	0.334159	0.0655257	0.109417
Negatius ampliat a distància 10. Validats amb distància 10	0.145667	0.874512	0.248802
Tots els positius. Negatius ampliat a distància 1. Vali- dats amb distància 10	0.738524	0.0270759	0.052231
Tots els positius. Negatius ampliat a distància 2. Vali- dats amb distància 10	0.670457	0.0299584	0.057346
Tots els positius. Negatius ampliat a distància 3. Vali- dats amb distància 10	0.567755	0.0351699	0.0662211
Tots els positius. Negatius ampliat a distància 5. Vali- dats amb distància 10	0.448267	0.0849888	0.142741
Tots els positius. Negatius ampliat a distància 10. Va- lidats amb distància 10	0.269577	0.89989	F-score: 0.414162

Taula 5.5: Taula de resultats amb negatius ampliat sobre els propis models.

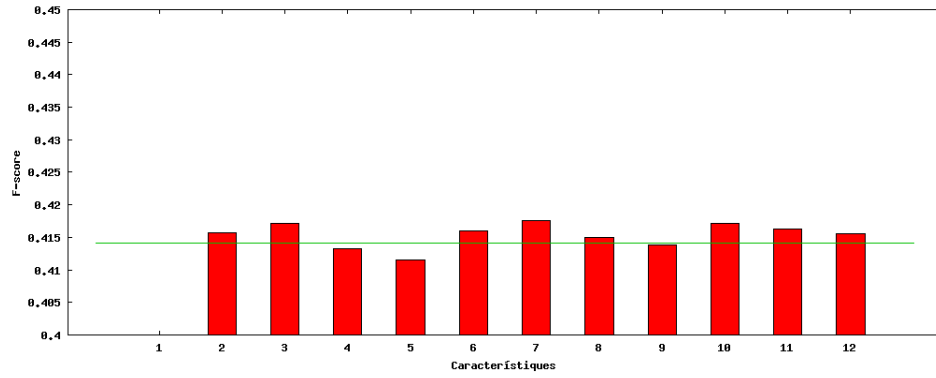


Figura 5.1: Variacions entrenant el model eliminant una a una les característiques.

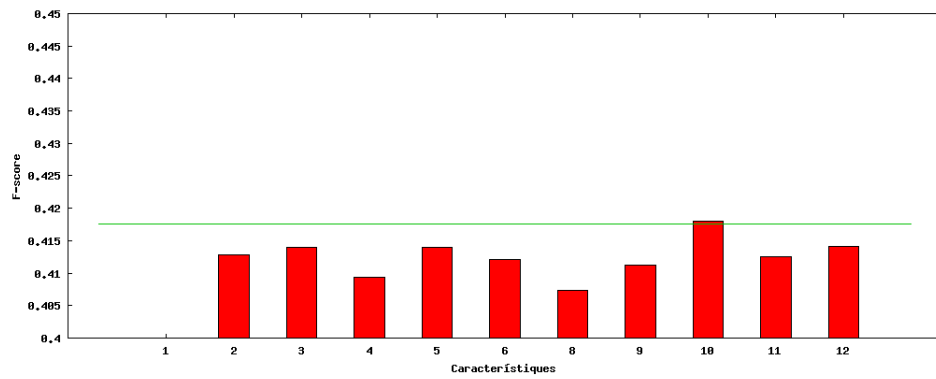


Figura 5.2: Variacions entrenant el model sense el *número* eliminant una a una les característiques.

Les característiques eliminades són les següents:

- 1 Distància
- 2 i-Pronoun
- 3 j-Pronoun
- 4 String Match
- 5 Definite Noun Phrase
- 6 Demonstrative Noun Phrase
- 7 Number Agreement
- 8 Gender Agreement
- 9 Semantic Class Agreement
- 10 Both Proper Names
- 11 Alias
- 12 Appositive

Per últim, com que hem vist que la característica que semblava indispensable era la distància, s'ha fet una nova prova entrenant un model amb només la característica de distància. Sorprenentment el resultat no varia gaire, inclús sembla que millora una mica. En concret:

Recall: 0.279775 Precision: 0.872824 F-score: 0.423143

Una de les primeres coses que podem concloure és que realment a part de la distància, la resta de característiques no aporten casi res en el millor dels casos, o simplement afegixen soroll que pot empitjora el resultat. Si comparem amb el millor resultat que teníem:

Recall: 0.271568 Precision: 0.90804 F-score: 0.417588

Es pot veure que només amb la distància tenim resultats molt similars; augmentem una mica el *Recall* i baixem la *Precision*, el *F-score* augmenta poc, però augmenta.

5.5 Corbes d'aprenentatge

Les corbes d'aprenentatge ens serveixen per veure com es comporta l'algorisme d'aprenentatge en front del nombre de regles que utilitzem per aprendre. A les figures 5.3 i 5.4 tenim les corbes del *F-score* utilitzant totes les característiques i només la distància respectivament. Es pot veure que són gràfiques força semblants, el que reforça la idea de que poc aporten les característiques apart de la distància.

A les figures 5.5 i 5.6, tenim les mateixes taules però sobre la *Precision*. Aquí si que podem veure que en el cas de tenir totes les característiques, la precision és una mica més alta. Al contrari que amb el recall en les figures 5.7 i 5.8, que sembla una mica millor si només treballem amb la distància. De totes maneres el fet de que no hi hagin diferències prou significatives reforça la idea de que la majoria d'informació per fer l'aprenentatge ve de la distància.

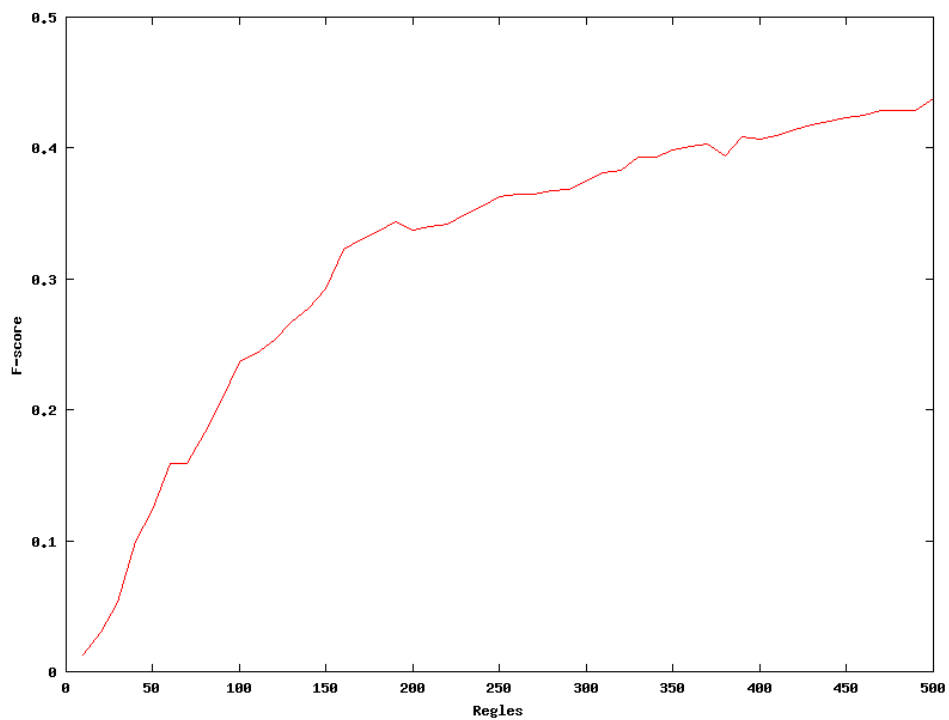


Figura 5.3: Corbes d'aprenentatge sobre F-score amb totes les característiques.

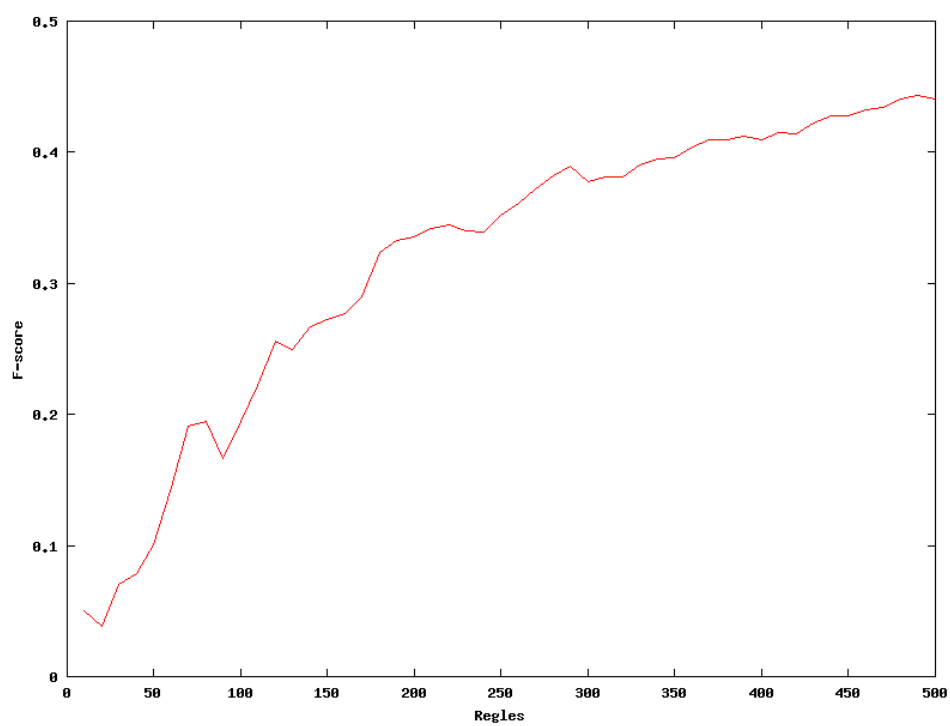


Figura 5.4: Corbes d'aprenentatge sobre F-score únicament amb la característica de distància.

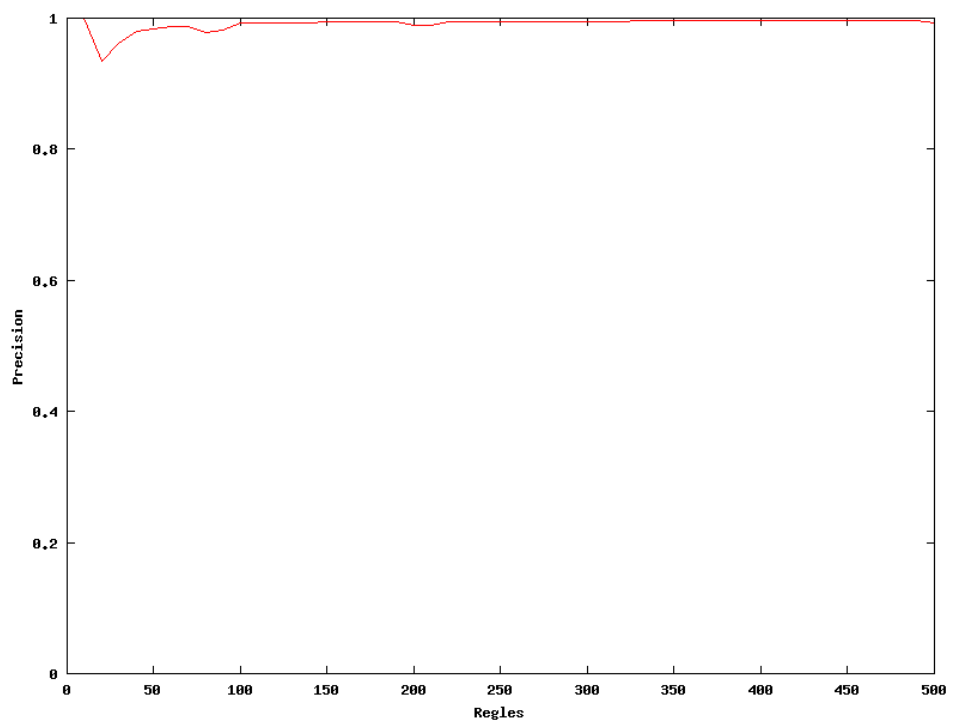


Figura 5.5: Corbes d'aprenentatge sobre Precision amb totes les característiques.

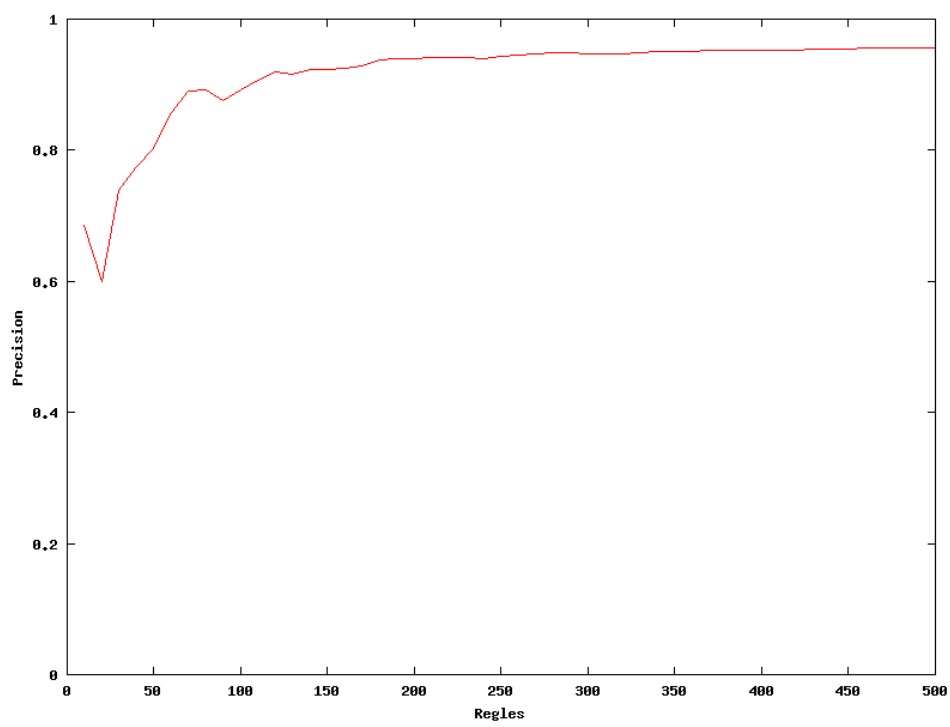


Figura 5.6: Corbes d'aprenentatge sobre Precision únicament amb la característica de distància.

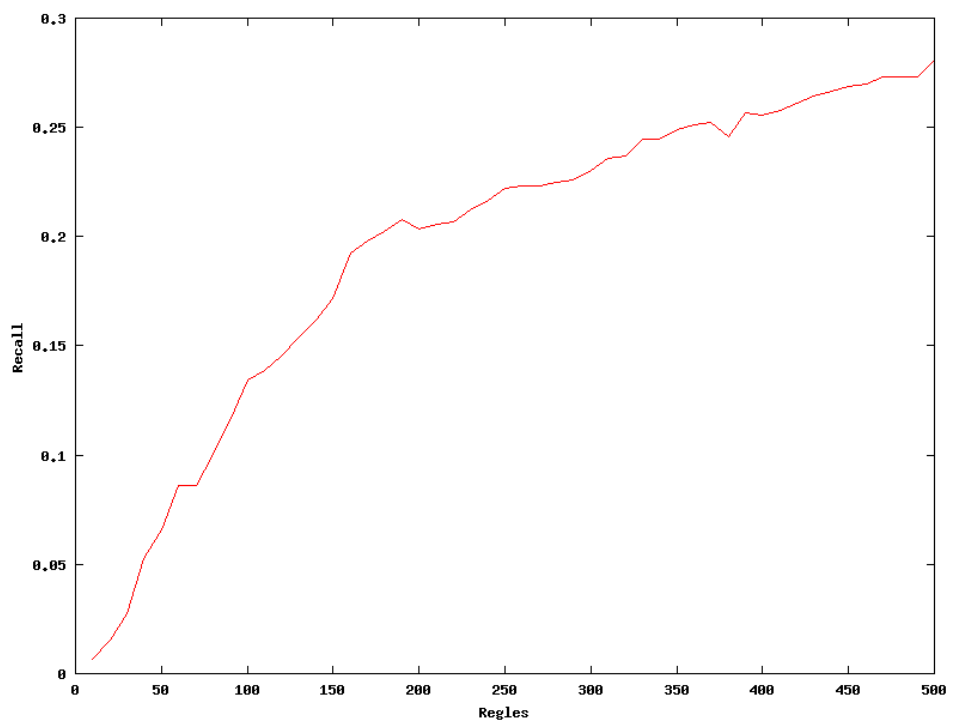


Figura 5.7: Corbes d'aprenentatge sobre Recall amb totes les característiques.

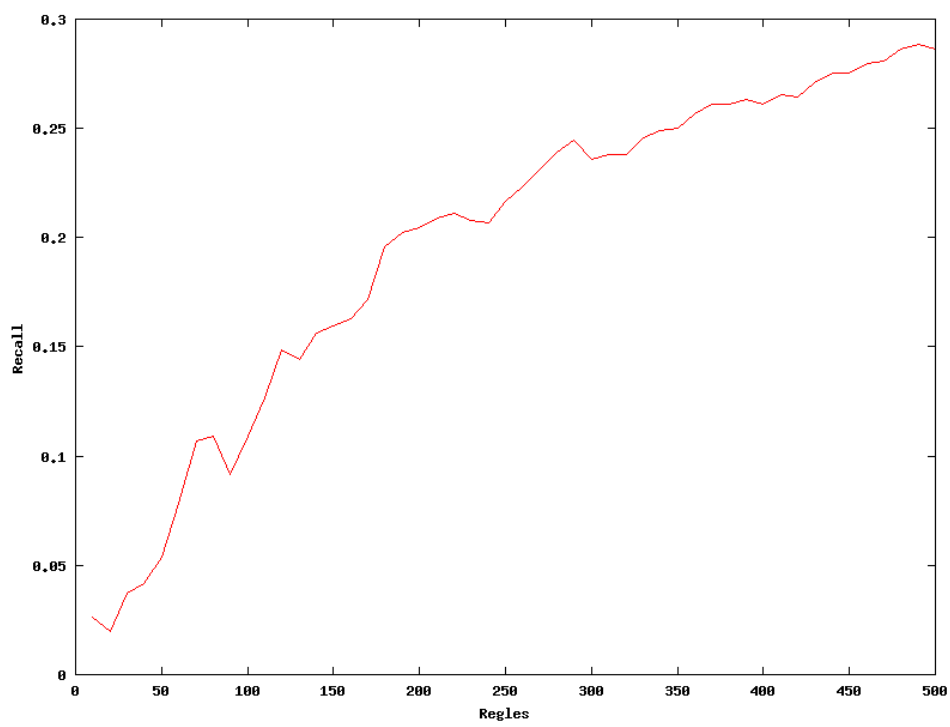


Figura 5.8: Corbes d'aprenentatge sobre Recall únicament amb la característica de distància.

Capítol 6

Planificació i costos

Aquest projecte consta de dues parts. La primera i més important és l'entrenament del model a partir del corpus *AnCora*, i la segona és la integració al *Freeling* com un mòdul més. L'entrenament del model, es pot dividir en quatre etapes (procés del corpus *AnCora*, extracció de característiques, entrenament de models i validació dels models).

6.1 Planificació inicial

En un primer moment, es va plantejar una planificació al llarg d'un any per les limitacions de temps de dedicació al projecte (unes 20 hores setmanals i entre 30 i 40 setmanes).

Etapa	Inici	Finalització
Procés del corpus <i>AnCora</i>	Gener 2008	Març 2008
Extracció de característiques	Març 2008	Maig 2008
Entrenament del model	Juny 2008	Juliol 2008
Validació del model	Juliol 2008	Agost 2008
Integració al <i>Freeling</i>	Setembre 2008	Octubre 2008
Conclusions i documentació	Novembre 2008	Desembre 2008

Taula 6.1: Planificació inicial.

6.2 Imprevistos i planificació real

Al llarg del desenvolupament del projecte, van sorgir varis imprevistos, que van fer que la planificació original es veies alterada considerablement.

1. **Procés del corpus *AnCora*.** El primer imprevist va ser veure que al corpus *AnCora* no estava marcat el *PoS tagging* de les paraules, i de cara a extreure les característiques era necessari. Es va tenir de processar en paral·lel i de forma sincronitzada el corpus *CESS* per extreure aquesta informació.
2. **Entrenament del model.** El problema trobat és que l'excés de exemples negatius en front dels positius feia que el model només aprenia a reconèixer exemples negatius, classificant-los tots com negatius. Això va provocar que s'hagués de revisar el procés del corpus, així com l'extracció de característiques.
3. **Validació del model.** Aquest va ser l'imprevist que més va afectar a la planificació. El problema venia de que al haver eliminat exemples negatius per tal de poder entrenar, el model es comportava bé amb aquest subconjunt d'exemples, però quan s'intentava validar amb un conjunt d'exemples amb tots els negatius, el resultat era molt dolent (veure el capítol 5).

Donat els problemes que es van trobar alhora de generar un conjunt adequat d'exemples positius i negatius, i del resultat d'entrenar aquests models, ha fet que la primera part del projecte (entrenar un model) s'hagi retrasat molt. Sobre tot perquè obligava a revisar la forma de generar els exemples (al procés del corpus), com a l'extracció de característiques. Finalment, degut a l'ajust de tots els mòduls, s'han acabat encavalcant en el temps. L'única etapa que ha durat el temps esperat, i fins hi tot menys ha sigut l'integració al *Freeling*, degut que al final la decisió de processar tots els *sn* simplificava molt l'algorisme, i a més és independent de l'entrenament del model. La resta d'etapes s'ha allargat el desenvolupament fins el final.

Etapa	Inici	Finalització
Procés del corpus <i>AnCora</i>	Gener 2008	Desembre 2008
Extracció de característiques	Març 2008	Desembre 2008
Entrenament del model	Juny 2008	Desembre 2008
Validació del model	Juliol 2008	Desembre 2008
Integració al <i>Freeling</i>	Novembre 2008	Novembre 2008
Conclusions i documentació	Novembre 2008	Gener 2008

Taula 6.2: Planificació final.

6.3 Costos

Per valorar el cost econòmic del projecte primer definim quins perfils necessitem:

Perfil	Preu/hora
Cap de projecte	30€
Analista	25€
Programador	20€

Taula 6.3: Perfils necessaris.

El cost en temps a acabat sent més del esperat. Si en un principi s'havia calculat un temps al voltant de 600 hores (20 hores a la setmana x 30 setmanes), ara hem d'afegir el temps extra utilitzat. Aquest temps han sigut 240 hores, és a dir, unes 12 setmanes, tot hi que no han sigut setmanes reals, sinó hores extres.

Aquestes 840 hores podem veure quin cost econòmic té assignant els rols:

Perfil	Preu/hora	Hores	Total
Cap de projecte	30€	168 (20%)	5040€
Analista	25€	252 (30%)	6300€
Programador	20€	420 (50%)	8400€
Total		840	19740€

Taula 6.4: Cost del projecte.

Capítol 7

Conclusions

La primera conclusió que podem treure, és que en el nostre sistema, la proposta de *Soon et al.* ([6]) no funciona bé. Hi han dos problemes importants que han impedit que el resultat final sigui l'esperat i que haurien de ser revisats en un futur treball.

7.1 Problemes i possibles solucions

7.1.1 Problema 1. Positiu vs Negatiu

Fent cas de l'algorisme de *Soon et al.* ([6]) generem un excés de exemples negatius, amb l'agreujant de que no generem negatius molt pròxims que després seran molt importants. Ja hem vist que canviant la forma de generar els negatius i augmentant el número de positius (generant totes les possibles combinacions) millorem el resultat. Tot hi que la idea general es manté, sobretot per extreure les característiques i per respondre amb un model ja entrenat.

Això pot ser degut en part a la forma en que esta etiquetat el corpus *Ancora*. Ja em vist que si fem al peu de la lletra la seva proposta el sistema és incapaç d'aprendre res, i que s'han tingut de fer variacions en la forma d'extreure els exemples per tenir un sistema que funcione. Al apèndix A es pot trobar un recull de casos del corpus *AnCora* que entren en conflicte amb la metodologia utilitzada.

7.1.2 Solució 1. Enriquir el corpus *AnCora*

Per veure com es podria millorar la generació d'exemples més enllà dels canvis fets, hauríem de canviar l'etiquetatge al corpus *AnCora* afegint positius que no estan actualment marcats. Com es pot veure en el següent exemple:

```
<de id="de_0" type1="da0ms0" type2="NE-org">
  <w>El</w>
  <w>grupo</w>
  <w>estatal</w>
  <de id="de_1" type1="np0000o" type2="NE-org">
    <w>Electricité_de_France</w>
    <de id="de_2" type1="np0000o" type2="NE-org">
      <corefLink anchor="de_0" id="84" type="pred"/>
      <w>-fpa-</w>
      <w>EDF</w>
      <w>-fpt-</w>
    </de>
  </de>
</de>
```

En aquest cas podem veure que el *de_2* està marcat com anàfora de *de_0*, i en canvi el *de_1* no està marcat. Seguint l'algorisme de *Soon et al.* ([6]), llavors tenim com a positiu *de_0-de_2* i com a negatiu *de_1-de_2*. Però això és incorrecte, ja que *de_1* i *de_2* estan fent referència a la mateixa entitat. És més, és un acrònim. Observant els fitxers de l'*AnCora*, es pot veure que aquestes formacions són força comuns.

Una forma d'intentar corregir aquest fals negatiu, ha sigut fent una regla *ad hoc* que elimina els negatius si el referent és intern a un altre DE que és alhora positiu. L'inconvenient és que amb aquesta regla també podem estar eliminant negatius reals molt propers (i per tant importants). Una vegada implementat, l'impacte en els resultats ha sigut menyspreable, i en alguns casos contraproductiu.

També hem de considerar que quan em ampliat la tria de negatius, també em inclòs com a negatiu *de_0-de_1*.

Quan agafem aquesta frase amb el *Freeling*, retorna el següent arbre:

```
S_ [
  sn_ [
    espec-ms_ [
      +j-ms_ [
        +(El e1 DA0MS0 -)
```

```

    ]
  ]
  +grup-nom-ms_[
    +n-ms_[
      +(grupo grupo NCMS000 -)
    ]
    s-a-ms_[
      +a-ms_[
        +(estatal estatal AQOCS0 -)
      ]
    ]
  ]
]
sn_[
  +grup-nom-ms_[
    +w-ms_[
      +(Electricité_de_France
        electricité_de_france NP00000 -)
    ]
  ]
]
sn_[
  (( ( Fpa -)
  +sn_[
    +grup-nom-ms_[
      +w-ms_[
        +(EDF edf NP00000 -)
      ]
    ]
  ]
  ( ) ) Fpt -)
]
F-term_[
  +( . . Fp -)
]
]

```

Si tenim en compte que el mòdul de coreferència del *Freeling* agafa com a candidats tots els *sn*, llavors podem veure clarament el problema. Els casos que es consultaran al model seran:

1. *El grupo estatal i Electricité_de_France*

2. *Electricité de France* i (*EDF*)
3. *El grupo estatal* i (*EDF*)

Per lògica hauríem d'esperar que les dues primeres fossin positives, i que la tercera no s'arribés a fer, degut a que la segona a tingut èxit. En canvi l'algorisme d'aprenentatge, només ha tingut com exemples positius la tercera, així que difícilment pot aprendre a respondre a les dues primeres. A més amb l'agreujant de què la primera surt com exemple negatiu, i la segona també depenent si decidim filtrar els casos ja esmentats.

Sembla clar que el problema és de manca d'etiquetatge al corpus *AnCora*, ja que hauria de ser clar quins *DE*'s són positius i quins són negatius. És molt difícil quantificar l'impacte d'aquest problema, degut a que hauríem de modificar el corpus, però tot indica que ha de ser prou important. Per un altra banda, la solució no sembla costosa, ja que augmentant els casos positius (marcant els que falten) ja aconseguiríem la nostra fita de generar un conjunt d'exemples més consistent.

7.1.3 Problema 2. Característiques que no aporten res

Una de les coses més sorprenents, és veure que l'única característica que al final ha sigut rellevant sigui la distància. Això sembla un clar indicador de que hi ha un problema greu. Potser que també s'estigui patint problema 7.1.1, ja que en el cas de *Soon et al.* ([6]), s'arribava a la conclusió que les característiques més rellevants eren la distància, la aposició i el tractament d'àlies. En canvi en aquest mòdul sembla que ni l'aposició ni l'àlies aporten gaire. Tal com podem veure a l'apèndix A, alguns dels problemes trobats al generar exemples incorrectes, són precisament aposicions i àlies (acrònims per ser més concrets), que generen exemples negatius incorrectes. És possible que aquesta sigui la raó per la que l'algorisme d'aprenentatge no acaba d'aprendre a resoldre aquests casos que per un altre banda haurien de ser de fàcil resolució.

Altres característiques que es podrien millorar en un treball futur:

1. **Number Agreement.** Ara només es tracta amb l'etiquetatge que aporta el *Freeling*. És podria millorar tractant els *sn* que representen un conjunt. Per exemple “*El jugador y su entrenador*“ seria plural. Altres casos que s'hauria de veure com resoldre són també els acrònims i les entitats nombrades. Per exemple *ONU* pot ser tant *Las naciones unidas* o *La organización de las naciones unidas*.
2. **Gender Agreement.** En el cas de les entitats nombrades, no tenim la

informació de gènere. Es podria intentar aportar aquesta informació. Per exemples amb els noms de les persones.

3. **Alias.** És podria substituir per un algorisme més sofisticat, com per exemple *Alias Assignment in Information Extraction* [13].

7.1.4 Problema 3. Extracció de *SN* al *Freeling*

Un possible problema és el poder garantir que des de el *Freeling* obtindrem els mateixos *exemples* que els que tenim marcats a l'*AnCora*, i així saber que tenim el mateix tipus d'exemples a l'entrenament i a les consultes. Sembla que aquesta condició no es dóna. En part per la pròpia naturalesa del corpus *AnCora* que el marcat de *DE*'s està en forma d'arbre tenint *DE*'s que inclouen d'altres i al mòdul del *Freeling* hem optat per agafar els *sn* que ens dóna marcats l'arbre de sintaxis.

7.1.5 Millora 1. Característiques

Una altra millora podria ser afegint noves característiques amb informació sintàctica. Sembla que com a mínim dins de la mateixa sentència, es podria fer servir per detectar negatius. Per exemple entre el que fa la funció de subjecte i un complement indirecte. També podria ser que afegís massa soroll i que acabés perjudicant en altres casos. Si més no seria interessant d'avaluar.

7.1.6 Millora 2. Entrenament

Un problema que sempre ha estat present és que tant en el corpus *AnCora* com en qualsevol text real, el nombre de parelles coreferents són una minoria en proporció a les parelles que no són coreferents. Això significa que un conjunt d'exemples si es vol que sigui representatiu respecte al tipus de problema tractat, ha de tenir molts més exemples negatius que positius. Llavors, com hem pogut comprovar això és un problema per un algorisme d'autoaprenentatge, ja que s'acostumen a comportar millor quan la proporció és més equilibrada, o com a mínim amb no tanta diferència. Una possible millora podria ser trobar algun algorisme d'aprenentatge diferent de l'*Adaboost* (*Shaphire & Freund* [7]) que tingués un millor comportament davant d'aquesta situació. Segurament si es provés un algorisme que també funcionés amb arbres de decisió seria més fàcil d'integrar amb el *Freeling*.

Apèndix A

Conflictes a AnCora

Part dels problemes que s'han trobat tenen el seu origen en el corpus *AnCora* i la forma en que es processa. Aquests problemes poden ser tant de manca d'etiquetatge a *AnCora* com de tractament deficient del propi corpus. Queda com una tasca futura solucionar aquests tipus d'exemples generats de forma incorrecta. Seguidament hi ha un petit recull de problemes trobats al processar *AnCora* que apareixen amb relativa freqüència.

- Un cas en que no hi ha cap relació amb el *de_7*, i que per tant generaria negatius falsos. L'acrònim *de_8* hauria de ser exemple positiu amb *de_7*.

```
<de id="de_6" type1="da0fs0" type2="NE-org">
  <w>la</w>
  <w>empresa</w>
  <w>mexicana</w>
  <de id="de_7" type1="np0000o" type2="NE-org">
    <w>Electricidad_Águila_de_Altamira</w>
    <de id="de_8" type1="np0000o" type2="NE-org">
      <corefLink anchor="de_6" id="85" type="pred"/>
      <w>-fpa-</w>
      <w>EAA</w>
      <w>-fpt-</w>
    </de>
  </de>
</de>
...
```

- Un cas d'estructura similar a l'anterior on si que és correcte marcar els exemples negatius de *de_1* amb *de_0* i *de_2*.

...

```

<de id="de_0" type1="di0ms0" type2="nne">
  <w>Un</w>
  <w>decenio</w>
  <w>de</w>
  <de id="de_1" type1="ncfp000">
    <w>cumbres</w>
    <w>iberoamericanas</w>
  </de>
  <w>,</w>
  <de id="de_2" type1="rel" type2="nne">
    <corefLink anchor="de_0" id="356" type="ident"/>
    <w>que</w>
  </de>
...

```

- En aquest cas, *de_4* i *de_3* també haurien d'estar marcats, així com *de_2* i *de_3*.

```

...
  <de id="de_2" type1="da0ms0+nccs000"
    type2="NE-pers">
    <w>atleta</w>
    <w>cubano</w>
    <de id="de_3" type1="np0000p"
      type2="NE-pers">
      <w>Javier_Sotomayor</w>
    </de>
  </de>
  <w>,</w>
</de>
<w>aseguró</w>
<w>que</w>
<de id="de_4" type1="da0ms0" type2="spec">
<corefLink anchor="de_2" id="72" type="ident"/>
  <w>el</w>
  <w>plusmarquista</w>
  <w>mundial</w>
  <w>de</w>
...

```

- En aquest exemple sembla que *de_33* i *de_34* haurien de ser coreferents (i estar marcats), perquè fan referència al mateix lloc. Sinó ens trobem en que *de_35* i *de_34* apareix com exemple negatiu, així com *de_33* i *de_34*.

```

...
<de id="de_33" type1="da0fs0" type2="NE-loc">
  <w>la</w>
  <w>localidad</w>
  <w>francesa</w>
  <w>de</w>
  <de id="de_34" type1="np00001">
    <w>Montauban</w>
    <w>,</w>
    <de id="de_35" type1="rel" type2="spec">
      <corefLink anchor="de_33" id="84"
        type="ident"/>
      <w>donde</w>
    </de>
  </de>
...

```

- Un altre cas amb un *de* al mig de dos que són coreferents i que ens portarà a generar exemples negatius que no haurien de ser.

```

...
<de id="de_40" type1="da0fs0" type2="NE-org">
  <w>la</w>
  <w>empresa</w>
  <de id="de_41" type1="np0000o" type2="NE-org">
    <w>Rubin</w>
  </de>
  <w>,</w>
  <de id="de_42" type1="rel" type2="spec">
    <corefLink anchor="de_40" id="295" type="ident"/>
    <w>que</w>
  </de>
...

```

- Un exemple més. No queda clar si els exemples negatius que es generarien són correctes o no.

```

...
<de id="de_15" type1="da0fp0" type2="NE-loc"
type3="nne">
  <w>las</w>
  <w>islas</w>
  <de id="de_16" type1="np00001" type2="NE-loc">
    <w>Malvinas</w>
  </de>

```

```

<w>,</w>
<de id="de_17" type1="rel" type2="spec">
<corefLink anchor="de_15" id="102" type="ident"/>
  <w>que</w>
</de>
...

```

- Aquest és un exemple no d'un problema, sinó de perquè no és bona la regla de no utilitzar els negatius si l'element es interior a un altre que si es coreferent. Tenim com exemples positius *de_13* i *de_14*, i *de_13* i *de_16*. Però per aprendre a resoldre el cas *de_13* i *de_16*, es important tenir l'exemple negatiu *de_15* i *de_16*.

```

...
<de id="de_13" type1="np0000p" type2="NE-pers">
  <w>Jack_Agrios</w>
  <de id="de_14" type1="ncms000" type2="NE-pers">
    <corefLink anchor="de_13" id="56" type="pred"/>
    <w>,</w>
    <w>presidente</w>
    <w>del</w>
    <de id="de_15" type1="da0ms0+np0000o" type2="NE-org">
      <w>Comité_Organizador_de_Edmonton_2001</w>
    </de>
    <w>,</w>
  </de>
  <de id="de_16" type1="rel" type2="spec">
    <corefLink anchor="de_13" id="57" type="ident"/>
    <w>quien</w>
  </de>
...

```

- Un altre exemple de negatiu fals.

```

...
<de id="de_25" type1="da0fs0" type2="NE-org">
<falseposLink anchor="de_6" id="332"/>
  <w>la</w>
  <w>compañía</w>
  <w>escandinava</w>
  <de id="de_26" type1="np0000o" type2="NE-org">
    <w>SAS</w>
  </de>
  <w>-Fpa-</w>

```

```
<de id="de_27" type1="rel" type2="nne">  
<corefLink anchor="de_25" id="245" type="ident"/>  
  <w>que</w>  
</de>
```

...

Apèndix B

Glossari

Adaboost. Algorisme d'aprenentatge automàtic. És una implementació concreta (*Shaphire & Freund [7]*) de la tècnica de *boosting*.

Anàfora. Relació de coreferència on el sintagma nominal que fa la primera menció a la entitat (nominal o entitat nombrada) apareix abans en el text que el coreferent.

Ancora. AnCora és un corpus del català i de l'espanyol amb diferents nivells d'anotació:

- categoria morfològica
- constituents i funcions sintàctiques
- estructura argumental i papers temàtics
- classe semàntica verbal
- sentits de Wordnet nominals
- entitats nombrades

Arbre de decisió. És un model de predicció utilitzat en l'àmbit de la intel·ligència artificial. Donat un conjunt de dades es construeix un arbre on cada node equival a una condició o pregunta. Cada fulla de l'arbre equival a una classe, i les condicions dels nodes que van des de la arrel fins a la fulla és una regla de classificació.

Boosting. Tècnica d'aprenentatge automàtic. Es tracta d'un sistema de combinació de classificadors elementals (*classifier ensembles*). Els classificadors elementals solen ser mini-arbres de decisió.

C4.5. Tipus d'algorisme d'aprenentatge basat en arbres de decisió.

C5. Tipus d'algorisme d'aprenentatge basat en arbres de decisió. És una evolució del C4.5

Catàfora. Relació de coreferència on el sintagma nominal que fa la primera menció a la entitat (nominal o entitat nombrada) apareix després del coreferent. La catàfora normalment és un recurs literari i poc freqüent.

CESS. CESS és un projecte l'objectiu del qual ha estat la creació de tres corpus, un per l'espanyol (CESS-ESP), un pel català (CESS-CAT) i un per l'euskera (CESS-EUS), de 500.000 paraules els dos primers i de 350.000 l'últim, etiquetats sintàcticament (amb constituents i funcions els corpus CESS-ESP i CESS-CAT i amb dependències el corpus CESS-EUS) i semàntica, amb els synsets nominals de WordNet.

Coreferència. Relació entre dues unitats lingüístiques que refereixen a la mateixa entitat del món real.

Crossvalidation. O validació creuada. És una pràctica estadística en la qual es divideix el conjunt de dades en subconjunts i de forma iterativa es van agafant tots per aprendre menys un que s'utilitza per validar el resultat.

DE. *Discourse Entities*, o entitats del discurs. Són les unitats lingüístiques que apareixen en el document, en *Ancora* venen marcades per el tag corresponent.

Fries. És una llibreria *open source* que ens permet convertir texts en llenguatge natural en vectors de característiques pensats per servir com a *input* d'un algorisme d'aprenentatge automàtic. A més, ens proporciona els tipus de dades per representar els documents.

Freeling. Llibreria d'eines lingüístiques. [9]

F-score. És la mitja harmònica ponderada de la *Precision* i el *Recall*. Quan la ponderació de la *Precision* i el *Recall* són iguals, llavors es calcula com: $F_1 = \frac{2*Precision*Recall}{Precision+Recall}$

Llenguatge natural. El llenguatge natural és el llenguatge parlat i/o escrit i/o signat per humans per propòsits generals de comunicació.

MUC. *Message Understanding Conferences* Són unes competicions sobre l'extracció d'informació. En concret la sisena i la setena (MUC-6 i MUC-7) incorporaven informació sobre entitats nombrades i coreferència.

NER. *Named entity recognition* Reconeixement d'entitats nombrades.

NEC. *Named entity classification* Classificació d'entitats nombrades.

NLP. *Natural Language processing* Processament del llenguatge natural.

Omlet. És una llibreria *open source* que ens ofereix eines orientades a l'aprenentatge automàtic.

POS tagging. *Part-of-speech tagging*, també anomenat etiquetat gramatical o desambiguació morfosintàctica. És el procés d'assignar (o etiquetar) a cada una de les paraules d'un text la seva categoria gramatical. Aquest procés es pot realitzar en base a la definició de la paraula o el context en que apareix, per exemple la seva relació amb les paraules veïnes en una frase, oració o en un paràgraf.

Precision. És la mesura que ens diu quan exacte és un sistema sobre els positius retornats. També es coneix com a *ràtio d'acceptació*. Es calcula com el número de positius marcats correctament dividit per el número de positius marcats.

Recall. O exhaustivitat. És la mesura que ens diu quan complet és un sistema amb els positius trobats. És la seva capacitat. Es calcula com el número de positius marcats correctament dividit per el número de positius esperats.

Recuperació de la informació (Information Retrieval) és la ciència de la cerca d'informació en documents, ja sigui a través de Internet, texts o dades d'altres característiques, de forma rellevant.

Referència És la relació entre una unitat lingüística i una entitat del món real.

Bibliografia

- [1] Leslie Lamport. *L^AT_EX – A Document Preparation System*. Addison-Wesley, second edition, Reading, MA, 1994.
- [2] Helmut Kopka and Patrick W. Daly. *A Guide to L^AT_EX*. Addison-Wesley, fourth edition, Boston, MA, 2004.
- [3] Michel Goossens, Frank Mittelbach, et al. *The L^AT_EX Companion*. Addison-Wesley, second edition, Boston, MA, 2004.
- [4] ŠOCHMAN, Jan. *Evaluation of the AdaBoost*. Center for Machine Perception, Prague, Czech Republic. <http://cmp.felk.cvut.cz>
- [5] BAEZA-YATES, Jan y RIBEIRO-NETO, B. *Modern Information Retrieval*. Addison Wesley, 1999
- [6] WEE MENG SOON , HWEE TOU NG , DANIEL CHUNG YONG LIM *A machine learning approach to coreference resolution of noun phrases*, Computational Linguistics, v.27 n.4, December 2001
- [7] YOAV FREUND AND ROBERT E. SCHAPIRE *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences, 55(1):119–139, August 1997.
- [8] XAVIER CARRERAS AND ISAAC CHAO AND LLUÍS PADRÓ AND MUN TSA PADRÓ *FreeLing: An Open-Source Suite of Language Analyzers*, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), 2004.
- [9] JORDI ATSERIAS AND BERNARDINO CASAS AND ELISABET COMELLES AND MERITXELL GONZÁLEZ AND LLUÍS PADRÓ AND MUN TSA PADRÓ *FreeLing 1.3: Syntactic and semantic services in an open-source NLP library*, Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genoa, Italy. May, 2006. <http://www.lsi.upc.edu/~nlp/freeling>
- [10] RECASENS, MARTA *Towards Coreference Resolution for Catalan and Spanish*. Master Thesis Universitat de Barcelona, 2008.

- [11] MARTÍ, M. A., MARIONA TAULÉ, LLUÍS MÀRQUEZ AND MANUEL BERTRAN *CESS-ECE: A Multilingual and Multilevel Annotated Corpus*, Pending to be published, 2007.
- [12] TAULÉ, M., M.A. MARTÍ, M. RECASENS *Ancora: Multilevel Annotated Corpora for Catalan and Spanish*. Proceedings of 6th International Conference on Language Resources and Evaluation. Marrakesh (Morocco), 2008.
- [13] EMILI SAPENA, LLUÍS PADRÓ AND JORDI TURMO *Alias Assignment in Information Extraction*, TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain, *Procesamiento del Lenguaje Natural*, no39 (2007), pp. 105-112.
- [14] JAVIER BÉJAR ALONSO *Aprendizaje*, Universitat Politècnica de Catalunya <http://www.lsi.upc.es/~bejar/apren/apren.html>.
- [15] *The XML C parser and toolkit of Gnome* <http://xmlsoft.org/html/index.html>.