UNIVERSITAT DE BARCELONA

Grau de Lingüística

Treball de Fi de Grau

Curs 2019-2020

Creació d'un corpus d'*entailment* en espanyol

Nom de l'estudiant: Patricia Grau Francitorra

Nom del tutor: Mariona Taulé Delor

Barcelona, 12 de juny de 2020

Resum

L'estudi de la inferència en el llenguatge natural i la seva detecció pot suposar un avenç important en les tecnologies del llenguatge. Per aquest motiu, s'han creat corpus per a tasca de *Natural Language Inference*, que estudien l'*entailment* i la contradicció, si bé han estat centrats en l'anglès. Aquest treball presenta la metodologia duta a terme per a la creació i anotació manual de frases d'un corpus d'*entailment* en espanyol. Especialment, s'ha descrit el procés de creació de frases inferides de textos a través d'uns criteris que n'asseguren la seva riquesa, que hi hagi diferents nivells de complexitat i eviten que hi hagi informació esbiaixada. S'han creat 940 hipòtesis inferides a partir de 470 frases inicials, que van ser extretes de 6 articles de la Viquipèdia. El corpus d'*entailment* en espanyol forma part d'un corpus més gran de *Natural Language Inference* que s'està duent a terme en el marc d'un projecte que desenvolupa el grup de recerca CLiC (Centre de Llenguatge i Computació) de la Universitat de Barcelona.

Paraules clau: corpus, *entailment*, anotació, *Natural Language Inference*, contradicció.

Abstract

The study of Natural Language Inference and its detection may suppose an important advance in language technology. For this reason, many corpora regarding entailment and contradiction have been created, even though most of them have been written for English. This work presents the methodology for the creation and annotation of a corpus of entailed sentences in Spanish made by humans. Especially, the process of creation of entailed sentences from texts through some criteria that ensure its richness, different levels of complexity and lack of bias. 940 hypotheses have been entailed from 470 texts, which were taken from 6 Wikipedia articles. The corpus of entailment in Spanish is part of a larger corpus about Natural Language Inference, which is being developed in the project of the research group CLiC (Centre de Llenguatge i Computació) of the University of Barcelona.

Keywords: corpus, entailment, annotation, Natural Language Inference, contradiction.

Agraïments

M'agradaria donar les gràcies a les professores Mariona Taulé i M. Antònia Martí, per donar-me l'oportunitat de treballar en la creació del corpus d'*entailment* i poder fer la meva petita aportació en aquesta recerca. En concret, gràcies per escoltar els meus dubtes i donar-me pautes i consell, sense els quals el resultat d'aquest treball seria molt diferent.

Voldria agrair especialment al doctorand Venelin Kovatchev, el responsable de la recerca de *Natural Language Inference* en la que s'insereix aquest corpus. Ell és qui va proposar la idea de crear un corpus en castellà i qui ens ha orientat durant tot el projecte.

Finalment, voldria donar les gràcies a totes les persones que han participat en la creació i anotació del corpus d'*entailment* i contradicció, i a tots els que ens han recolzat en aquest procés. Sense vosaltres, aquest treball no hauria estat possible.

Índex

Taules

1. Introducció

En el marc del Processament del Llenguatge Natural (PLN), l'estudi de la inferència de coneixement ha avançat considerablement en els darrers anys per millorar una sèrie de tasques relacionades amb la comprensió del llenguatge natural. És el que en anglès es coneix amb el nom de *Natural Language Inference* (NLI). L'objectiu de la NLI és determinar si la relació semàntica entre dues expressions lingüístiques (oracions o paràgrafs) és una relació d'implicació (1a), una relació de contradicció (1b) o bé una relació indeterminada (1c). Vegem l'exemple (1):

1. T - *Alfons el Magnànim va fundar l'actual Universitat de Barcelona l'any 1450*.
   1a. H - *La Universitat de Barcelona existeix des del segle XV*.
   1b. H - *María Josefa Amalia de Sajonia, casada amb Ferran VII, va fundar la Universitat de Barcelona*.
   1c. H – *L'edifici històric de la Universitat de Barcelona va ser declarat monument historicoartístic nacional*.

(1a) i (1b) són les frases (hipòtesis, H) inferides a partir de la frase inicial (1). La diferència rau en què (1a) és una inferència certa ("entailment") i la (1b) és una inferència falsa ("contradiction"). La inferència és informació que no és explícita en la frase inicial i que només es pot inferir a partir del coneixement del món, és a dir a partir de fets del món implícits i ben sabuts per tothom. El darrer exemple (1c) és un cas de frase no relacionada semànticament amb la frase inicial (o text, T) (1).

Per tant, el tractament de les relacions d'implicació i contradicció se centren en la factualitat, en si els fets que s'expressen són certs, falsos o indeterminats; en la inferència, és a dir, en informació implícita; i en el coneixement del món, a partit de l'ús de fets sobre el món que són generalment coneguts.

L'estudi del tractament automàtic de les relacions d'inferència, com són la implicació i la contradicció, és necessari per a millorar diferents aplicacions de PLN, com els sistemes de cerca de respostes, l'extracció d'informació o el resum automàtic. Per a poder dur a terme aquesta tasca, cal disposar de corpus anotats, preferiblement per humans. Aquests corpus serveixen no només per estudiar la implicació, sinó també per entrenar i avaluar sistemes basats en aprenentatge automàtic.

En aquest marc de treball s'està desenvolupant un projecte liderat per en Venelin Kovatchev[1], l'objectiu principal del qual és el desenvolupament d'una metodologia per a la creació i anotació d'un corpus d'*entailment* i contradicció en espanyol. La contradicció s'explica en el Treball de Fi de Grau de la Yauheniya Verkhavodkina.

L'objectiu d'aquest treball és la descripció de la metodologia aplicada en la creació i anotació d'un corpus d'*entailment* en espanyol. Aquest treball està especialment centrat en la creació d'hipòtesis relacionades per *entailment*, en la definició dels criteris que es van utilitzar en la creació de les hipòtesis i en les dificultats que ens hem trobat en aquest procés. El corpus s'ha dut a terme per estudiants de lingüística de la Universitat de Barcelona i ha estat supervisat per experts en la matèria.

[1] El projecte s'està duent a terme en el Centre de Llenguatge i Computació (CLiC) de la Universitat de Barcelona.

2. Marc teòric

La variabilitat del llenguatge natural i la seva relació amb l'ambigüitat són unes de les qüestions que més dificultat impliquen per a les aplicacions de Processament del Llenguatge Natural (PLN). En concret, "un fenomen fonamental del llenguatge natural és la variabilitat de l'expressió semàntica, on el mateix significat pot ser expressat o inferit de diferents textos" (Dagan, Dolan, Magnini, & Roth, 2009, p. ii).

Una de les maneres d'estudiar la variació semàntica és a través de l'*entailment* o implicació textual. La definició clàssica d'implicació parteix de la lògica formal i, en l'àmbit de la lingüística, es considera part de la semàntica. La implicació o *entailment* es pot definir com una relació entre un parell de proposicions de manera que la veritat de la segona frase necessàriament es deriva de la veritat de la primera (Crystal, 1998). A aquesta definició, Akmajian (1997) afegeix que la falsedat de la segona proposició garanteix la falsedat de la primera.

Es tracta, però, d'una definició clàssica difícilment adaptable a aplicacions de PLN. Dagan, Glickman i Magnini (2006, p. 1) en prenen una de més pràctica o "informal":

> L'*entailment* o implicació textual es defineix com la relació direccional entre dos textos, denotats per T, el "text" del que es treuen les implicacions, i H, les "hipòtesis" implicades. Diem que T implica H si el significat de H pot ser inferit pel significat de T.

Aquesta és la definició que proposen en *The PASCAL Recognising Textual Entailment Challenge*[2]. El *RTE Challenge* és una competició on diversos grups proposen aplicacions per reconèixer, donats dos fragments de text, si el significat d'un text es pot inferir (o implicar) del significat de l'altre. Va començar al 2005 per avaluar i comparar sistemes que tracten l'*entailment* en un mateix entorn.

Des del seu començament, els *RTE Challenges* han promogut la recerca en el reconeixement de l'*entailment*, ja que creen un marc comú per estudiar les inferències semàntiques que es necessiten en moltes aplicacions de PLN. Algunes de les aplicacions de PLN que es poden beneficiar de l'estudi de la variació semàntica són: els sistemes de cerca de respostes (*Question Answering*, QA), l'extracció d'informació (*information extraction*, IE) i el resum automàtic (*Summarisation*, SUM). En el primer cas, un sistema de QA podria extreure les respostes implicades d'un text. Per exemple, donada la pregunta

[2] http://www.pascal-network.org/Challenges/RTE/

“Qui va escriure la novel·la ‘Orgull i prejudici’?”, la resposta “Jane Austen” es podria inferir del text següent: “La ironia que Jane Austen utilitza repetidament en Orgull i prejudici dona un to còmic als seus personatges”, d'on se'n deriva a través de la relació d'implicació que Jane Austen va escriure 'Orgull i prejudici’. La recuperació d’informació també seguiria un procés similar, ja que l’objectiu dels buscadors és trobar la resposta a una pregunta en diferents documents. En el cas del resum automàtic, el sistema hauria de poder determinar que una informació està recollida en una part del text per no repetir-la.

L’*entailment* forma part de la *Natural Language Inference* (NLI), que és la tasca de determinar si una hipòtesi és certa (*entailment*), falsa (contradicció) o indeterminada o desconeguda (neutral), donada una premissa o un text inicial.

En la creació d’aquest corpus se segueix la definició d’*entailment* proposada per Dagan *et al.* (2009), que és la definició generalment més acceptada en tasques de NLI. Per poder avaluar l’*entailment* en diferents aplicacions es necessita disposar de corpus anotats amb aquesta informació. A continuació, farem un breu resum dels corpus disponibles anotats amb *entailment*. La majoria s’han fet en anglès o amb traduccions a partir de l’anglès, i n’hi ha pocs en altres llengües.

2.1 Corpus anotats amb *entailment*

En aquesta secció es descriuen breument els diferents corpus que estudien la *Natural Language Inference* i en particular, l’*entailment*, en diferents llengües.

Recognising Textual Entailment (RTE) Challenge

El procés de creació de corpus en els *RTE challenges* comença per la creació d’un *dataset* amb parells de text-hipòtesis (parells T-H). Aquests parells s’anoten manualment com a “True”, si les hipòtesis es poden inferir del text, o “False”, quan T no implica H.

Els textos que es van utilitzar al primer *RTE challenge* (RTE1) es van extreure tant de fonts externes (*datasets* preexistents) com d’internet, especialment de notícies, i estaven escrits en anglès. Es van escollir els parells de T-H tenint en compte diferents aplicacions de PLN: recuperació i extracció d’informació, comparació de documents, comprensió lectora, cerca de respostes, traducció automàtica i paràfrasi.

El procés d’anotació constava de tres parts. En primer lloc, els anotadors que van crear els parells T-H els anotaven com a cert o fals. Després, els exemples eren revisats per uns segons anotadors, que feien les revisions sense tenir cap tipus de context de les frases.

Les frases en què no hi havia acord van ser eliminades del corpus (un 20% del total). Finalment, un dels organitzadors del *RTE1 challenge* va revisar els exemples i va eliminar un 13% addicional perquè eren exemples que podrien crear controvèrsia. Així, totes les frases restants es van considerar el *Gold Standard* d'avaluació. Aquestes es van dividir en 567 exemples per al entrenament dels sistemes (*training set*) i 800 exemples de test que contenien anotacions "True"/"False" de forma equitativa. Per tant, els parells T-H van ser anotats per dos anotadors diferents com a mínim i, la majoria, per tres.

En les següents edicions del *RTE challenge* els criteris d'anotació van canviar. En la quarta edició (*RTE4*), l'anotació dels parells de frases es feia mitjançant tres categories diferents: "entailment", "contradiction" (contradicció) i "unknown" (desconegut). En algunes tasques, les contradiccions i desconegudes es van classificar en un gran grup anomenat "no entailment".

La definició de contradicció que agafen Dagan *et al*. (2009) per a la tasca d'implicació textual prové de De Marneffe, Rafferty and Manning (2008, p. 1040). Segons aquests autors, la contradicció ocorre entre dues frases quan "és molt improbable o impossible que les dues siguin veritat al mateix temps". Giampicollo *et al.* (2008, p. 4) consideren que un parell de frases es pot etiquetar com a "unknown" quan "la veritat d'H no es pot determinar partint de T".

La detecció de la contradicció és, per tant, una altra relació de significat molt útil que es pot implementar en aplicacions de PLN, juntament amb l'*entailment*. Dagan *et al*. (2009, p. xiii) comenten, però, que la detecció de la contradicció sembla ser una tasca més difícil que reconèixer la implicació, ja que requereixen d'inferències més profundes, avaluant la correferència i la construcció de models.

Semantic Evaluation (SemEval)

Els *PASCAL RTE Challenge* és un dels marcs de referència per a l'estudi de l'*entailment* i la contradicció, però no n'és l'únic. El *SemEval* (*Semantic Evaluation*) és una altra competició o avaluació de sistemes computacionals d'anàlisi semàntica que busquen estudiar la naturalesa del significat en el llenguatge. Els primers anys es van centrar en la desambiguació automàtica de paraules (WSD*, Word Sense Disambiguation*) i durant la seva llarga trajectòria es van tractar diferents aplicacions del llenguatge centrades en la semàntica. No va ser fins al 2010 (*SemEval2010*) que aquestes conferències incorporen

el tractament de l'*entailment* en les seves tasques, en concret, en la tasca 12: *Parser Evaluation using Textual Entailments*.

Posteriorment, *SemEval* va dur a terme una tasca que tractava l'*entailment* en diferents llengües: el *Cross-lingual Textual Entailment for Content Synchronization* (SemEval 2012, tasca 8). L'objectiu de la tasca era promoure la recerca de la inferència semàntica en textos escrits en diferents llengües, per identificar relacions d'*entailment* multi-direccionals ("forward entailment", "backward entailment", "bidirectional entailment", "no entailment").

Per crear un corpus multilingüe, Negri *et al.*(2012) van crear un corpus monolingüe en anglès a partir de dades extretes manualment de la Wikipedia i Wikinews. Cada text es va modificar per obtenir una hipòtesi i els parells text-hipòtesi es van etiquetar seguint les relacions de "forward entailment", "backward entailment", "bidirectional entailment" o "no entailment"). Després, els textos (T) en anglès es van traduir manualment a quatre llengües diferents: espanyol, alemany, italià i francès, a través de traductors experts. D'aquesta manera es va crear un corpus multilingüe en què els parells T-H podien ser espanyol-anglès, alemany-anglès, italià-anglès i francès-anglès.

Negri (2010) ja havia treballat anteriorment amb la traducció de corpus per estudiar *entailment*, si bé havia utilitzat una estratègia diferent: crear traduccions a través de *crowdsourcing* mitjançant l'eina *Mechanical Turk* de què disposa *Amazon*.

Una altra estratègia és la que van seguir Zhao *et al.* (2013) amb l'ús d'eines de traducció automàtica per poder treballar amb parells multilingües. En aquest cas, parells T-H en diferents llengües (espanyol-anglès, alemany-anglès, italià-anglès i francès-anglès) es van traduir a través del traductor de Google per tenir-los en la mateixa llengua. Altres autors han utilitzat la traducció automàtica per estudiar l'*entailment*, com Miquel Esplà-Gomis *et. al* (2012) en la mateixa tasca de SemEval 2012.

Standford Natural Language Inference corpus (SNLI)

El SNLI és un corpus creat al 2015 en anglès per estudiar la N*atural Language Inference* (NLI) que conté 570.000 parells de text-hipòtesi. Es tracta, per tant, d'un corpus de gran mida, exactament de dos ordres de magnitud més gran que els altres recursos preexistents per estudiar NLI.

Els autors d'aquest corpus defineixen la NLI com "l'ús de l'*entailment* i la contradicció en sistemes computacionals, conceptes semàntics que són centrals en tots els aspectes del significat del llenguatge natural" (Bowman, Angeli, Potts, & Manning, 2015, p. 1).

El SNLI va ser creat amb *crowdsourcing* a través de l'*Amazon Mechanical Turk*. Els contribuïdors del *Mechanical Turk* havien de crear hipòtesis a partir de títols de fotografies que s'havien recollit prèviament en una altra tasca de *crowdsourcing*. En aquest cas, com en la quarta edició del RTE *challenge*, les hipòtesis no es basaven només en la detecció o creació de frases d'implicació, sinó que també incloïen contradiccions i frases neutrals. Així, de cada text (títol de fotografia) i amb la informació de coneixement del món s'havien de crear tres hipòtesis: una que fos una descripció "definitivament vertadera" de la imatge (*entailment*), una que "podria ser vertadera" (neutral) i una que fos "definitivament una descripció falsa" de la fotografia (contradicció).

Per assegurar la qualitat del corpus, un 10% d'aquestes hipòtesis van seguir un procés d'anotació: trenta treballadors del *Mechanical Turk* van anotar parells de T-H amb les etiquetes "entailment*"*, "contradiction" o "neutral". Cada parella de T-H es va presentar a quatre anotadors diferents i, juntament amb l'etiqueta posada per l'autor de les frases, es van obtenir cinc etiquetes per cada parell T-H. L'etiqueta "gold" es donava quan tres dels autors coincidien en la seva anotació i només es van deixar de banda un 2% dels casos.

Tot i que la Kappa de Fleiss és conservadora, Bowman *et al.* (2015, p. 635) consideren que "la taxa d'acord global és extremadament alta, cosa que suggereix que el corpus és d'una qualitat suficientment alta com per posar una tasca desafiadora però realista a l'ús de sistemes basats en aprenentatge automàtic".

SPARTE

Un corpus que va ser creat en espanyol sense cap tipus de traducció és SPARTE (Peñas, Rodrigo, & Verdejo, 2006). Aquest corpus es va basar en els corpus de QA usats al CLEF (*Cross-Language Evaluation Forum*) en les últimes tres edicions. Per poder treballar amb els parells T-H es van generar les hipòtesis de forma automàtica a partir de les preguntes dels corpus anteriors. Els autors donen un exemple molt clar (Peñas *et al.*, 2006, p. 3): de la pregunta "Quina és la capital de Croàcia?" es genera la hipòtesi "La capital de Croàcia és <resposta/>". Les possibles respostes s'extrauen de diferents documents i s'etiqueten com a "correcta", "no confirmada", "inexacta" o "incorrecta" segons diferents criteris.

Cross-lingual Natural Language Inference corpus (XNLI)

El Cross-lingual Natural Language Inference corpus o XNLI (Conneau et al., 2020) és un corpus creat per estudiar la NLI en àmbits multilingües. Aquest corpus consisteix en 112.500 parells de text-hipòtesi anotats manualment amb les etiquetes "entailment", "contradiction" i "neutral" en 15 llengües: anglès, francès, espanyol, alemany, grec, búlgar, rus, turc, àrab, vietnamita, thai, xinès, hindi, swahili i urdú. L'objectiu és avaluar NLI en diferents llengües, de manera que un model s'entrena en una llengua i es fa el test en unes altres.

Aquest corpus multilingüe parteix d'un corpus en anglès que forma part del *Multi-Genre Natural Language Inference Corpus* (MultiNLI) (Williams, Nangia, & Bowman, 2018). A partir de 2500 frases que provenen de deu fonts diferents, els treballadors de *crowdsourcing* van crear tres hipòtesis diferents (*entailment*, contradicció i neutral). Aquestes van ser anotades posteriorment per quatre anotadors fins a obtenir cinc etiquetes per cada parell T-H, on es va arribar a un acord de tres o més anotadors en el 93% dels casos. Finalment, aquests 7500 textos i hipòtesis van ser traduïts en quinze llengües a través de la plataforma *One Hour Translation*.

El corpus que s'ha desenvolupat i que presentem en aquest treball, "Corpus d'*entailment* en espanyol", és un corpus nou creat a partir d'informació nova. És a dir, els textos no provenen de cap corpus previ, sinó que s'han extret d'internet, en concret d'articles de la Viquipèdia. És un corpus creat i anotat per humans directament en espanyol, sense cap tipus de traducció.

3. Metodologia

En aquest apartat es descriu la metodologia per a la creació d'un corpus d'*entailment* en castellà que no està basat en corpus previs. Seguint les propostes de RTE4 i altres, aquest corpus no se centra només en l'*entailment*, sinó que incorpora també la contradicció i neutralitat (parells T-H on no hi ha ni implicació ni contradicció). El corpus de contradicció es recull al treball final de grau de la Yauheniya Verkhavodkina, una de les moltes persones que han estat implicades en la creació del corpus.

Aquest treball s'inclou dins d'una recerca sobre NLI, el responsable de la qual és en Venelin Kovatchev. L'objectiu principal és la creació d'un corpus d'*entailment* i contradicció. Per a la creació del corpus s'han realitzat dues tasques fonamentals: la creació d'hipòtesis i l'anotació d'aquestes. La meva participació s'ha centrat en la primera tasca, la creació d'hipòtesis, tot i que també vaig treballar en l'anotació del corpus.

Per a la creació d'aquest corpus s'han dut a terme les etapes que es detallen a continuació: la selecció dels textos (secció 3.1), la creació de les implicacions o *entailments* (secció 3.2) i l'anotació dels parells T-H (secció 3.3).

3.1 Selecció del text (T)

Per poder crear parells de frases que es puguin utilitzar en tasques d'*entailment* es necessiten text (T) i hipòtesis (H) derivades del text. La selecció dels textos es va fer a partir d'articles de la Viquipèdia[3] en castellà que tractaven sis temes diferents. Els diferents articles es van escollir per ser informació de caràcter general, coneguts per la majoria de la població, de caràcter descriptiu i amb informació factual. El seu caràcter objectiu i descriptiu facilita la creació de les hipòtesis.

Es van escollir sis articles de la Viquipèdia que tracten temes diversos de caràcter general: el pintor Pablo Picasso; el descobridor d'Amèrica Cristòbal Colón; un esdeveniment esportiu mundial, els Jocs Olímpics (JJOO); els videojocs; i dues unions polítiques, la Unió Europea (UE) i la Unió de Repúbliques Socialistes Soviètiques (Unió Soviètica, URSS).

[3] Les frases que van formar els textos pel corpus d'*entailment* i contradicció es van extreure de la Viquipèdia entre el 21 de gener i el 18 de març de 2020.

De cada article de la Viquipèdia es van agafar tantes frases com paràgrafs tingués l'article. D'aquesta manera, es podrien provar les aplicacions d'*entailment* i contradicció de dues maneres: amb parells de frases (T - H) o amb parells paràgraf (T) – frase (H). En el segon cas es podrà observar si el context ajuda als sistemes a identificar la implicació i contradicció.

Les frases que provenien de la Viquipèdia havien de tenir una mida específica: entre 15 i 45 tokens. Aquesta llargada mínima de 15 tokens limita els textos a frases prou llargues, on hi ha més informació i d'on es poden extreure hipòtesis més fàcilment. Així, els *entailments* són més rics, variats, llargs i complexos. Si prenem dues frases de l'article de Pablo Picasso, podem veure la diferència en la creació d'hipòtesis per frases de diferents llargades. Vegem l'exemple (2):

2. T- *Picasso superó el examen de ingreso en la Escuela de la Lonja* (frase inventada, 12 tokens).

D'aquest text podríem obtenir algunes hipòtesis, com les que trobem en els exemples (2a-2c):

2a. H - *Picasso aprobó el examen para entrar a la Escuela de la Lonja.*
2b. H - *Picasso fue a la Escuela de la Lonja*.
2c. H -*Se debe hacer un examen de ingreso para entrar en la Escuela de la Lonja*.

Si agafem, en canvi, una frase més llarga que conté més informació, podem extreure un major nombre d'hipòtesis, més variades i complexes. Un exemple de frase llarga és l'exemple següent (3):

3. T - *Estudiante brillante y precoz, Picasso superó en un solo día, a la edad de catorce años, el examen de ingreso en la Escuela de la Lonja, y se le permitió saltarse las dos primeras clases* (frase real del corpus, 35 tokens).

Vegem les possibles hipòtesis (3a) i (3b) implicades del text anterior:

3a. H - *Picasso destacaba tanto como estudiante que pudo saltarse la dos primeras clases de la Escuela de la Lonja.*

3b. H - *El joven Picasso consiguió entrar sin problemas en la Escuela de la Lonja, ya que superó el examen de ingreso en un solo día.*

Es pot observar que les hipòtesis creades en el segon cas són més llargues i més riques sintàcticament. Per tant, queda clar que la llargada dels textos dels quals es creen les hipòtesis és un factor que pot fer variar la qualitat i riquesa del corpus.

Per garantir la comprensió de les frases, es van substituir els pronoms que trobàvem en els T per les referències apropiades. D'aquesta manera es podien agafar frases aïllades amb tota la informació necessària. Una de les frases que trobem en l'article de la Unió Europea és l'exemple següent (4):

4. T- *Este tratado fundador buscaba aproximar vencedores y vencidos europeos al seno de una Europa que a medio plazo pudiese tomar su destino en sus manos, haciéndose independiente de entidades exteriores.*

Sense tenir un context, no podem saber a quin tractat s'està fent referència, de forma que crear les hipòtesis és més difícil. Per aquest motiu, es van substituir manualment tots els pronoms que apareixien en els textos per les referencies apropiades. En aquest cas, vegem l'exemple (5):

5. T - *El Tratado que institucionaliza la Comunidad Europea del Carbón y del Acero (CECA) buscaba aproximar vencedores y vencidos europeos al seno de una Europa que a medio plazo pudiese tomar su destino en sus manos, haciéndose independiente de entidades exteriores.*

Dels sis articles de la Viquipèdia que es van triar, es van escollir 470 frases com a text (T) per crear hipòtesis tant d'*entailment* com de contradicció, seguint els criteris mencionats prèviament. Com els articles de la Viquipèdia no tenen les mateixes característiques, entre aquestes la llargada total de l'article o la mida de les frases, es van obtenir quantitats de textos diferents de cada article.

L'article del qual es van extreure més textos és l'article de videojocs, d'on es van obtenir 107 textos. És a dir, que un 22,77% del total de textos tracten sobre els videojocs. L'article del qual es van obtenir menys textos és el de Cristòbal Colón, d'on se'n van treure 59 frases. Aquestes dades es recullen a la Taula (1):

Taula 1. Textos per a crear hipòtesis: article d'origen, quantitat i percentatge dels textos que representen

Article de la Viquipèdia	Textos/paràgrafs	Percentatge
Picasso	82	17,45%
Colón	59	12,55%
Videojocs	107	22,77%
Jocs Olímpics	73	15,53%
Unió Europea	68	14,47%
Unió Soviètica	81	17,23%
Total	**470**	**100%**

3.2 Creació d'hipòtesis (H)

A partir de cadascuna de les frases (T) extretes dels articles, es van crear manualment dues frases d'*entailment* o hipòtesis (H). Aquesta tasca es va realitzar en dues fases. La primera fase va servir per poder establir els criteris que se seguirien en la creació de les hipòtesis, que es van aplicar en la fase 2. Per establir els criteris, es van utilitzar els articles de Picasso i de Colón, que posteriorment van ser revisats perquè s'adherissin a la nova manera de crear hipòtesis.

Aquesta part del corpus l'han creat estudiants de lingüística de la Universitat de Barcelona, amb la supervisió d'experts. La creació d'hipòtesis, per tant, ha estat una tasca totalment controlada en el seu procés. Els parells de text-hipòtesi creats en aquesta fase formen el *Gold Standard* a partir del qual es crearan nous parells de T-H, que posteriorment s'hauran d'anotar (tercera etapa, secció 3.3).

Durant la primera fase de creació d'hipòtesis es va observar que cada lingüista utilitzava una sèrie de recursos diferents per crear les hipòtesis. Per tal d'augmentar la variació dels textos i evitar biaixos, es va repartir la tasca de creació d'hipòtesis d'*entailment* i contradicció. D'aquesta manera, els lingüistes que han treballat en aquest corpus han creat hipòtesis tant d'*entailment* com de contradicció.

A continuació, es presenten els criteris que es van utilitzar per a la creació d'hipòtesis en la fase 2. Aquests criteris garanteixen la riquesa del corpus, que hi hagi diferents nivells de complexitat i eviten que hi hagi informació esbiaixada.

<u>Negació</u>

Un dels primers criteris que es van seguir en la creació de les hipòtesis, tant d'*entailment* com de contradicció, va ser incloure la negació en la creació de les frases. Sempre que

fos possible, una de les frases havia de contenir negació, ja fos en l'oració principal o en una oració subordinada.

L'objectiu d'escriure una frase afirmativa i una altra negativa era disminuir el biaix per a les aplicacions de PLN que utilitzessin els corpus d'*entailment* i contradicció. La manera immediata i més fàcil de crear una contradicció és afegir-hi un element negatiu. Per tant, per evitar que la negació es prengués com una característica única de la contradicció en l'aprenentatge automàtic, aproximadament la meitat de les hipòtesis contenien negació.

A la primera fase de la creació d'hipòtesis, totes les frases negatives contenien o bé negació simple o bé negació morfològica. En la negació simple, l'expressió de la negació es duu a terme amb una partícula que pot ser un adverbi (com "no", "jamás" o "nunca"), un pronom ("nadie" o "nada") o una preposició ("sin"), mentre que en la negació morfològica apareixen afixos negatius ("a-", "in-" o "des-") en un mot. També vam trobar alguns casos de negació complexa, tot i que no molts, en l'ús d'expressions tals com "hasta que... no", "antes de... no" i "no... sino que".

Després de la revisió de les hipòtesis creades amb els fitxers de Picasso i Colón, vam veure que hi havia una estratègia que no s'havia utilitzat: la negació lèxica, a través de paraules i sintagmes com "ausencia de", "exclusión", "fallar", "negar", etc. Així vam arribar als dos primers criteris:

Criteri 1. Crear una hipòtesi afirmativa i una que contingui negació, sempre que sigui possible.

Criteri 2. Utilitzar totes les eines disponibles per crear la negació: negació morfològica (a través de prefixos com "des-", "in-", "a-"), negació lèxica (amb paraules com "fracaso", "falsa", "ignorancia" o "falta de") i partícules negatives ("sin", "nadie", "jamás").

Vegem uns exemples d'hipòtesis negatives, extretes dels articles de la Unió Europea (UE) i dels Jocs Olímpics (JJOO), on cadascuna de les hipòtesis mostra una estratègia de creació de negació diferent.

En l'exemple (6a) veiem una hipòtesi negativa d'un text de la Unió Europea. Aquesta hipòtesi conté la partícula negativa "no" i el sintagma "a pesar de".

6. T - E*l inglés como lengua materna es la tercera más hablada (13%), sin embargo, es la primera por el número total de hablantes (51%), seguida por el francés.*

6a. H - *El inglés **no** es la lengua materna más hablada del mundo, **a pesar de** ser la lengua que más personas hablan.*

El següent exemple (7) és un parell T-H sobre la Unió Europea, on la hipòtesi (7a) té com a subjecte el pronom negatiu "nadie":

7. T - *En cuanto a jurisdicción sobre el Mar, un país de la Unión Europea, Francia (con más de 11 millones de km²), tiene la más extensa Zona Económica Exclusiva del mundo.*

   7a. H - ***Nadie** supera a Francia en quilómetros cuadrados de jurisdicción sobre el mar, un país que dispone de más de once millones de km².*

El darrer exemple d'hipòtesi negativa inferida d'un text de la Unió Europea és l'exemple (8a), on trobem la preposició negativa "sin":

8. T - *Estas agencias han contribuido de manera significativa al funcionamiento efectivo de la UE, gracias a su especialización en áreas determinadas de la arquitectura comunitaria*.

   8a. H - *La Unión Europea no funcionaría de forma tan efectiva **sin** estas agencias y su trabajo en la arquitectura comunitaria.*

Els exemples següents (9-11) són parells de text-hipòtesi que han sorgit de l'article de la Viquipèdia sobre els Jocs Olímpics. L'oració que trobem en l'exemple (9a) és un exemple d'hipòtesi que conté negació morfològica, a través dels prefixos "des-" i "in-":

9. T - *Este efecto positivo se inicia en los años previos a la celebración y puede persistir durante varios años después, aunque no de forma permanen*te.

   9a. H - *El efecto positivo que surge tras la celebración **des**aparece **in**evitablemente.*

Vegem una hipòtesi que conté negació complexa (10a):

10. T - *Se celebraron eventos de carreras, pentatlón —consistente en eventos de salto de longitud, lanzamiento de disco (discóbolo), jabalina, carrera pedestre y lucha (boxeo, lucha libre, pancracio) y eventos ecuestres.*

    10a. H - *En el pentatlón **no** se realizaban actividades acuáticas, **sino que** todos los eventos se hacían en tierra.*

Finalment, l'exemple (11) és un parell de text-hipòtesi que mostra l'ús de negació semàntica a través del verb "carecer".

11. T - *El COI permite la formación de comités olímpicos nacionales que representan a naciones, sin verse obligadas a cumplir con estrictos requisitos relacionados con la soberanía política como lo demandan otras organizaciones internacionales.*
    11a. H - *Las personas que forman parte de los comités olímpicos nacionales* ***carecen*** *de obligaciones o requisitos de soberanía política.*

Llargada de les hipòtesis

A la primera fase de creació d'hipòtesis es va observar que la majoria d'H creades per *entailment* eren frases curtes, mentre que les hipòtesis de contradicció solien ser més llargues. Per evitar que la mida de les frases fossin un factor determinant en l'aprenentatge automàtic, és a dir, que les frases més curtes es relacionessin amb la implicació i les frases més llargues, amb la contradicció, vam decidir buscar variació en la llargada de les hipòtesis. Per aquest motiu, un dels criteris que vam establir va ser:

Criteri 3. Crear hipòtesis de diferent llargada, ja siguin d'*entailment* com de contradicció.

Vegem un exemple d'hipòtesis de mida diferent en un text de la Unió Soviètica; l'exemple (12). La primera té 12 tokens, mentre que la segona en té més del doble (25 tokens):

12. T - *En junio de 1941, durante la Segunda Guerra Mundial, la Alemania nazi junto a sus aliados invadió la Unión Soviética, un país con el que había firmado un pacto de no agresión llamado luego Pacto Ribbentrop-Mólotov.*
    12a. H - *La Unión Soviética fue invadida por Alemania durante la Segunda Guerra Mundial.*
    12b. H - *Antes de la invasión, Alemania y la Unión Soviética firmaron un pacto de no agresión durante la Segunda Guerra Mundial que fue llamado Pacto Ribbentrop-Mólotov.*

Complexitat i variació

Un altre criteri que està relacionat amb l'anterior és la complexitat. Com en un principi la majoria de les hipòtesis d'*entailment* creades eren molt curtes i simples, en la segona fase

vam voler crear frases més llargues i més complexes. Per aquest motiu, vam crear hipòtesis que continguessin oracions coordinades i subordinades.

Els següents parells de T-H creades a partir de l'article de Videojocs són exemples de frases amb coordinació (ja sigui a nivell sintagmàtic o a nivell oracional) i subordinació:

13. T - *La industria de los videojuegos es el sector económico involucrado en el desarrollo, la distribución, la mercadotecnia, la venta de videojuegos y del hardware asociado.*
    13a. H - *El desarrollo, distribución y venta del hardware asociado a los videojuegos también pertenece a esta industria.*
14. T - *Para el cliente, el alquiler de un videojuego puede ser un paso previo a la compra del mismo: si lo prueban y les gusta, lo compran.*
    14a. H - *Algunos clientes deciden comprar un videojuego solo tras haberlo probado, y por eso los alquilan.*

En la primera hipòtesi (13a), extreta del primer parell T-H, podem veure coordinació de substantius dins del subjecte ("desarrollo, distribución y venta"). En la següent hipòtesi (14a), en canvi, la coordinació es fa entre oracions: "Algunos clientes deciden comprar un videojuego solo tras haberlo probado" + "por eso los alquilan". També en la segona hipòtesi (14a) trobem una oració subordinada adverbial ("solo tras haberlo probado") en la primera part de la coordinació.

Els següents exemples, que tracten sobre la Unió Soviètica, també contenen hipòtesis amb coordinació d'oracions, en l'exemple (15a), i una oració subordinada de relatiu, en l'exemple (16a):

15. T - *Durante este período, la Unión Soviética continuó avanzando científica y tecnológicamente, lo que le permitió lanzar el primer satélite artificial Sputnik 1 y conseguir la hazaña de llevar por primera vez un ser vivo al espacio exterior: la perra Laika.*
    15a. H - *El primer satélite artificial se llamaba Sputnik 1 y se puso en órbita gracias al trabajo de los rusos.*
16. T - *Después de la política económica del comunismo de guerra llevada a cabo durante la Guerra Civil, el Gobierno soviético permitió que algunas empresas privadas coexistieran con la industria nacionalizada durante los años 1920.*

16a. H - *La política económica que caracterizó la Guerra Civil soviética fue el comunismo de guerra.*

Per evitar que les oracions del text i de les hipòtesis fossin molt semblants tant en la forma sintàctica com en la lèxica, es van fer una sèrie de transformacions sintàctico-semàntiques. Algunes de les oracions es van passar a passiva reflexa, es va canviar l'ordre de les paraules i algunes d'aquestes paraules es van substituir per sinònims, antònims o hiperònims.

Variar la forma de T i H és una altra mesura per intentar reduir el biaix al mínim en l'aprenentatge automàtic. Si les paraules que hi apareixen són molt similars, és més fàcil per a un sistema determinar que hi ha algun tipus de relació entre els parells de text a analitzar. L'ús de sinònims, antònims, hiperònims o hipònims enriqueixen el sistema i redueixen la possibilitat de que s'agafi la similitud de paraules com un tret identificador. Una eina útil va ser el diccionari de sinònims i antònims online (Naranjo, 2017).

El següent exemple (17), extret de l'article de Picasso, mostra una hipòtesi que conté una estructura passiva reflexa i una oració subordinada substantiva:

17. T - *Apollinaire fue arrestado bajo sospecha de haber robado la Mona Lisa en el Louvre, y ser parte de una banda de ladrones internacional*.

17a. H - ***Se cree que*** *Apollinaire era un ladrón que no trabajaba solo*.

En aquestes dues oracions dels exemples (18) i (18a) hi ha hagut un canvi en l'ordre de les paraules i una substitució de la paraula "pensamiento" pel sinònim "ideología".

18. T - *En este ambiente Picasso entró en contacto con el pensamiento anarquista, implantado en Barcelona.*

18a. H - *Barcelona era un centro de ideología anarquista en época de Picasso.*

En el següent parell text-hipòtesi (19) i (19a) sobre els Jocs Olímpics podem veure una estructura sintàctica diferent en T i H on hi apareix subordinació, ja que la hipòtesi comença per una oració subordinada substantiva que fa de subjecte:

19. T - *Los funcionarios griegos y el público en general estaban entusiasmados con la experiencia de albergar unos Juegos Olímpicos.*

19a. H - *Que los juegos se celebraran en Grecia causó alegría para la población griega.*

Un exemple de la hiperonímia és el següent parell de frases (20 i 20a) sobre la vida de Colón, en què "la Niña" i "la Pinta" s'han substituït per "las carabelas". Es tracta d'un hiperònim perquè la "Niña" i la "Pinta" són dos exemples d'aquest tipus d'embarcació:

20. T - *Finalmente, el 15 de marzo la Niña arribó al puerto de Palos, con pocas horas de diferencia respecto a la Pinta.*

20a. H - *A mediados de marzo, dos de las carabelas de Colón llegaron al puerto de Palos.*

Els següents criteris en la creació dels parells text-hipòtesi són els següents:

Criteri 4. Crear hipòtesis de diferent complexitat sintàctica: frases simples, coordinades i subordinades.

Criteri 5. Canviar l'ordre en què apareixen les paraules en els parells T-H (si és possible).

Criteri 6. Utilitzar paraules diferents en el text i les hipòtesis a través de l'ús de sinònims, antònims i hiperònims.

Correferència

Una altra estratègia per augmentar la variació entre textos i hipòtesis és utilitzar termes correferents. Crystal (1998, p. 116) defineix la correferència com un terme per referir-se a dos constituents d'una oració que tenen la mateixa referència. En aquest cas, es pren una visió més àmplia: quan tractem la correferència, trobem dos constituents que tenen la mateixa referència, però no cal que siguin de la mateixa oració.

Vegem els exemples (21) i (22), dos exemples de parells T-H extrets del corpus d'*entailment* de Picasso:

21. T - *En 1923* ***Picasso*** *continuó con el tema del arlequín; realizó varios retratos de Jacint Salvadó disfrazado de arlequín en un estilo menos monumental y más lírico.*

21a. H - *El arlequín era un tema recurrente para* ***el pintor cubista malagueño***.

22. T - *El 26 de abril de 1937 se produjo el bombardeo de Guernica por parte de la Legión Cóndor alemana a petición de* ***Franco***.

22a. H - ***El dictador español*** *mandó bombardear Guernica a finales de abril de 1937.*

En el primer exemple (21) es pot veure que "Picasso" i "el pintor cubista malagueño" són elements correferents, ja que els dos sintagmes designen la mateixa persona. El mateix ocorre en l'exemple (22), on ho són "Franco" i "el dictador español". Els elements correferents poden ser també pronoms (com "Colón" i "él") o poden fer referència a altres elements no personals, com llocs. Vegem-ho en el següent exemple (23):

23. T- *En este ambiente Picasso entró en contacto con el pensamiento anarquista, implantado en* ***Barcelona****.*

    23a. H- *Picasso no se mantuvo alejado de la corriente anarquista, que predominaba en* ***la capital catalana****.*

Els articles que van formar els textos del corpus feien referència a dues persones, Picasso i Colón; dos tipus de joc, videojocs i Jocs Olímpics; i dues unions polítiques, la Unió Europea i la Unió Soviètica. En cadascun d'aquest parell d'articles trobem elements que poden ser referits de la mateixa manera. En el cas de Picasso i Colón, trobem que "él" pot fer referència tant al pintor de Màlaga com al descobridor d'Amèrica. En el cas dels videojocs i dels Jocs Olímpics, "los juegos" és un sintagma que pot referir-se tant a uns com als altres, i en el cas de la Unió Europea i la URSS, passa el mateix amb "la Unión". Al tenir en compte elements correferents dels diferents corpus es disminueixen els biaixos en les futures aplicacions de detecció d'*entailment* que utilitzin aquest corpus.

De manera que el darrer criteri lingüístic per poder escriure hipòtesis és:

Criteri 7. Utilitzar termes correferents en algunes de les hipòtesis, per obtenir variació.

<u>Coneixement del món</u>

A l'hora de crear hipòtesis s'ha d'utilitzar el coneixement del món que tenim com a parlants per a poder fer inferència; s'ha de conèixer i utilitzar informació implícita sobre el món però coneguda per tothom. El coneixement del món s'expressa de diferents maneres, que poden ser més específiques o generals. Un exemple de coneixement del món específic és que Picasso era espanyol, ja que és una dada que gran quantitat de la població coneix. El coneixement del món general, en canvi, té a veure amb veritats universals. Per exemple, que una persona no pot estar en dos llocs alhora, així que si "a principis de 1920 Picasso era a París", com ens informa una frase del corpus, sabem que

en aquella època no era a Espanya ni a cap altre lloc que no fos la capital francesa. Per tant, un criteri que s'utilitza en la creació d'hipòtesis és:

Criteri 8. Es poden utilitzar fets coneguts en la creació d'hipòtesis (coneixement del món).

A continuació es mostren tots els criteris que es van decidir en la primera fase per a la creació de les hipòtesis del corpus.

Criteris per a la creació d'hipòtesis

Criteri 1. Crear una hipòtesi afirmativa i una de negativa, sempre que sigui possible.
Criteri 2. Utilitzar totes les eines disponibles per crear la negació: negació morfològica, negació lèxica i partícules negatives.
Criteri 3. Crear hipòtesis de diferent llargada, ja siguin d'*entailment* com de contradicció.
Criteri 4. Crear hipòtesis de diferent complexitat sintàctica: frases simples, coordinades i subordinades.
Criteri 5. Canviar l'ordre en què apareixen les paraules en els parells T-H (si és possible).
Criteri 6. Utilitzar paraules diferents en el text i les hipòtesis a través de l'ús de sinònims, antònims i hiperònims.
Criteri 7. Utilitzar termes correferents en algunes de les hipòtesis, per obtenir variació.
Criteri 8. Es poden utilitzar fets coneguts en la creació d'hipòtesis (coneixement del món).

Tot i seguir aquests criteris, la creació d'hipòtesis és un procés d'una certa complexitat, si es volen tenir en compte totes les variables necessàries, en especial, la variació entre les hipòtesis. En el següent apartat es recullen algunes de les dificultats que poden sorgir en la creació d'hipòtesis d'*entailment*.

Dificultats en la creació d'hipòtesis

Un dels factors que cal tenir en compte en la creació d'hipòtesis d'*entailment* és evitar crear el mateix tipus de frase. És molt fàcil crear hipòtesis molt similars, ja sigui en l'estructura sintàctica, en la manera d'afegir negació a les oracions o en els elements que la composen. Per aquest motiu es van idear els criteris (veure l'apartat anterior); amb

l'objectiu que el corpus sigui el més variat possible i evitar que hi hagi informació esbiaixada.

En alguns casos, però, crear hipòtesis diferents és més difícil. Per exemple, alguns dels textos a partir dels quals es creen les hipòtesis contenen negació, de manera que crear una frase afirmativa podria portar a que aquesta fos molt semblant sintàcticament al text o que fos molt simple. Vegem l'exemple (24), que tracta sobre la Unió Europea:

24. T - *Con la entrada en vigor del Tratado de Lisboa, los símbolos de la UE como la bandera, el lema, el himno o el Día de Europa no son jurídicamente vinculantes, aunque todos ellos se encuentran en uso.*

24a. H- *La bandera de la Unión Europea no es jurídicamente vinculante, así como tampoco lo son su lema o su himno.*

24b. H - *La bandera de la UE se usa en la actualidad pese al Tratado de Lisboa, en el que se explica que no es jurídicamente vinculante.*

En aquest exemple (24) trobem que el text conté la partícula negativa "no" i que les hipòtesis que s'infereixen d'aquest text també contenen negació. Es podrien haver creat altres exemples que no continguessin negació, vegeu exemples (24c-24f), però aquests són menys complexos, llargs o són afirmacions molt generals (24e). És cert que hi ha moltes més possibles hipòtesis que es podrien inferir d'aquest text, però no sempre són evidents per a la persona que les està escrivint.

24c. H - *La Unión Europea tiene bandera, lema e himno; signos que aún se mantienen.*

24d. H - *El Día de Europa sigue celebrándose.*

24e. H - *El Día de Europa existe.*

24f. H - *Uno de los tratados de la Unión Europea es el Tratado de Lisboa.*

Un altre element que pot dificultar la creació d'hipòtesis és el tema sobre el qual s'infereixen. Per exemple, el text de l'exemple (24) prové d'un article de la Unió Europea, que pot resultar difícil per a persones a les que no els interessa la política o que no tenen el castellà com a llengua primera, ja que contenen moltes paraules especialitzades.

Tot i a aquesta relativa dificultat, els textos factuals, és a dir, que contenen fets, faciliten la creació d'hipòtesis siguin del tema que siguin. L'exemple següent (25) mostra un text inventat que expressa opinió i unes hipòtesis que es poden inferir de T:

25. T - *Creo que te equivocas, porque lo que me dices no tiene ni pies ni cabeza.*

25a. H - *Tú no llevas la razón.*

25b. H - *Dices cosas que no tienen sentido.*

25c. H - *Estamos manteniendo una conversación.*

Com es pot veure, les hipòtesis que s'han extret són bastant curtes (25a-25b) o donen informació molt general (25c). Com en l'exemple anterior a aquest (24), les hipòtesis podrien ser molt variades, aquí només se'n mencionen unes quantes sense arribar a fer paràfrasis. Si bé és cert que moltes oracions que contenen opinió també poden contenir fets reals, en general és més fàcil partir d'oracions que no expressen opinió a l'hora de crear hipòtesis d'*entailment*.

Informació estadística del corpus d'*entailment*

Per cada un dels texts de cada article de la Viquipèdia es van crear dues hipòtesis, en les que s'intentava que hi hagués un balanç entre hipòtesis afirmatives i hipòtesis que continguessin negació. Com s'ha explicat en l'apartat de la selecció del text (vegeu apartat 3.1), els textos (T) dels quals s'inferirien les hipòtesis havien de seguir uns criteris: tenir entre 15 i 45 tokens, i provenir de paràgrafs diferents. D'aquesta manera, es van extreure quantitats de textos diferents de cada article, tal i com es veu a la Taula (2):

Taula 2. Quantitat de textos extrets de cada article de la Viquipèdia i hipòtesis d'*entailment* inferides dels textos

Article Viquipèdia	Text (T)	Hipòtesis (H)
Picasso	82	164
Colón	59	118
Videojocs	107	214
Jocs Olímpics	73	146
Unió Europea	68	136
Unió Soviètica	81	162
Total	**470**	**940**

Per tant, tant el corpus d'*entailment* com el de contradicció compten amb 940 hipòtesis, que provenen de 470 textos de 6 articles diferents. L'article de la Viquipèdia del qual s'ha extret més frases és l'article sobre videojocs, i l'article del que se n'han extret menys és el de Cristòbal Colón. A continuació, vegem a la Taula (3) les dades estadístiques sobre les hipòtesis del corpus d'*entailment*.

Taula 3. Nombre de tokens de les hipòtesis d'*entailment*

Article Viquipèdia	Mínim	Màxim	Mitjana
Picasso	5	34	16,01
Colón	7	36	18,26
Videojocs	6	29	15,63
Jocs Olímpics	8	33	17,58
Unió Europea	6	35	17,7
Unió Soviètica	5	33	16,57
Mitjana total			**16,96**

A la columna "Mínim" apareixen el nombre de tokens de les frases més curtes de cada article. Així, podem veure que les hipòtesis més curtes s'han inferit de textos que provenen dels articles de Pablo Picasso i de la Unió Soviètica. A la columna "Màxim" trobem el contrari; el nombre de tokens de les frases més llargues del corpus. Les hipòtesis més llargues, en aquest cas, sorgeixen de l'article de Cristòbal Colón. La darrera columna, "Mitjana", ens informa de la llargada mitjana de les hipòtesis creades de cada article. La mitjana més petita coincideix amb un dels corpus amb la hipòtesi més curta o "mínim" (Pablo Picasso) i la mitjana més gran, amb el corpus que té la frase més llarga (Cristòbal Colón). A la darrera fila de la taula, on es llegeix "Mitjana total" podem veure la mitjana aritmètica de la llargada de totes les hipòtesis del corpus d'*entailment*. En general, les hipòtesis creades tenen entre 16 i 17 tokens.

Creació d'hipòtesis per *crowdsourcing*

El conjunt de parells T-H creats en els apartats anteriors serviran de *Gold Standard* perquè han estat elaborats per anotadors experts que han participat en l'elaboració del corpus, en la selecció del text i en els criteris. La creació d'hipòtesis s'ha dut a terme amb més temps i ha estat supervisada per experts.

Ara bé, un corpus de 1880 parells de text-hipòtesis (940 d'*entailment* i 940 de contradicció) és insuficient per poder entrenar un sistema d'aprenentatge automàtic. És per això que, per tal d'augmentar el nombre de parells T-H i en conseqüència, la mida del corpus, s'han seguit dues estratègies.

La primera consisteix en la creació de noves hipòtesis, aplicant els mateixos criteris establerts, amb la diferència que aquest cop les hipòtesis no són creades per experts. Les hipòtesis les creen alumnes de 1r i 4t de lingüística i de màster, voluntaris, que realitzen la tasca de creació en una sessió única d'una hora de durada. Aquest tipus d'anotació

s'assembla a la anotació per *crowdsourcing*, ja que tenen un temps limitat per fer-la, seguint unes directrius, i no estan supervisats.

Aquesta és una tasca que encara s'està realitzant, per tal de tenir el major nombre de parells text-hipòtesi. La segona estratègia és la que s'explica a continuació, a l'apartat 3.3.

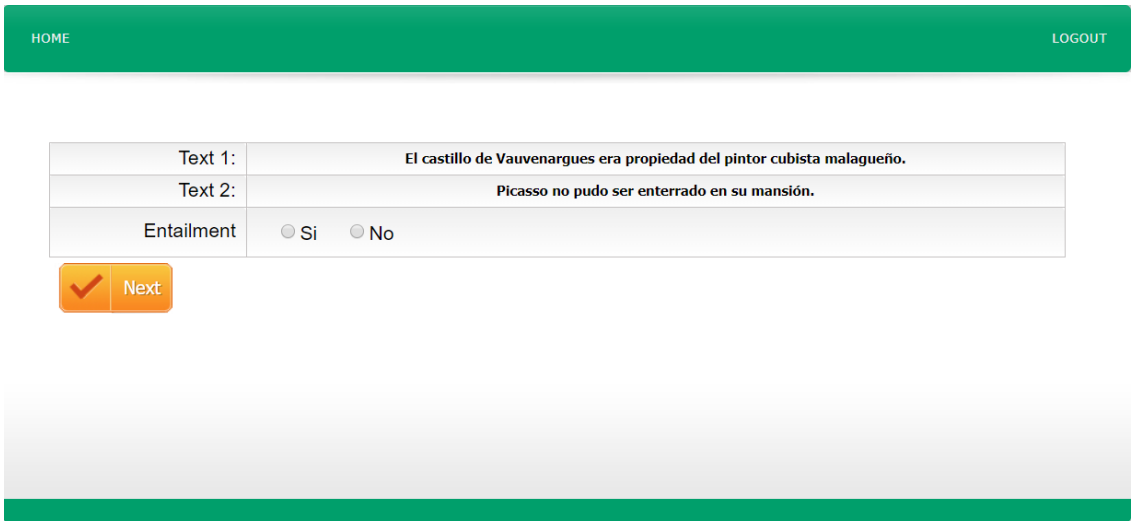
3.3 Anotació dels parells T-H

Una altra estratègia per augmentar el nombre de parells text-hipòtesi és combinar les oracions creades anteriorment entre elles i anotar quin tipus de relació tenen: "entailment", "contradicció" o "neutral".

L'anotació de parells d'oracions la duen a terme majoritàriament usuaris no experts. Les persones que van treballar en la creació d'hipòtesis també hi han contribuït, tot i que no és el seu objectiu principal. És una tasca on hi ha molta participació, ja que també hi han col·laborat els alumnes de lingüística de 1r i 4t curs i alumnes de màster; i encara s'està duent a terme.

L'anotació dels parells T-H és una tasca en la qual, donades dues frases, els anotadors han de dir si aquestes frases tenen una relació d'*entailment* o contradicció. Per fer-ho, s'utilitza una interfície creada especialment per aquesta tasca. La Imatge (1) és un exemple de la interfície que utilitzen els anotadors, en què apareixen dos textos (Text 1 i Text 2) i els anotadors han d'anotar, simplement, si el Text 1 implica o no el Text 2.

Imatge 1: Exemple de la interfície d'*entailment*



Es va crear un apartat per l'anotació d'*entailment* i un altre per l'anotació de contradicció; la Imatge (1) correspon al primer.

Les instruccions que reben els anotadors són les següents:

- En la detecció d'*entailment*: la tasca és determinar que, "si el Text 1 és cert, el Text 2 és definitivament cert."
- La direcció de l'*entailment* és sempre de Text 1 a Text 2.
- En la detecció de contradicció: la tasca és determinar que "Text 1 i Text 2 no poden ser certs alhora".

Es va insistir en la direcció de l'*entailment*. La implicació és una relació de significat en què hi ha una direcció. Si el Text 1 implica el Text 2, no necessàriament el Text 2 implica el T1. Vegem-ne uns exemples:

26. T - *Juan utiliza el metro cada día para ir a trabajar.*
    26a. H - *El metro existe.*
27. T - *El metro existe.*
    27a. H? - *Juan utiliza el metro cada día para ir a trabajar.*
28. T - *Juan utiliza el metro cada día para ir a trabajar.*
    28a. H - *Juan tiene trabajo*
29. T - *Juan tiene trabajo*
    29a. H? - *Juan utiliza el metro cada día para ir a trabajar.*

En l'exemple (26), podem veure que la hipòtesi s'infereix del text: si en Joan utilitza el metro per anar a la feina, és que el metro existeix. Però en el cas de l'exemple següent (27), podem veure que no és compleix l'altra direcció: que el metro existeixi no és una condició necessària perquè en Joan vagi a treballar amb aquest transport. El mateix ocorre en els exemples (28) i (29), ja que la relació d'*entailment* es dona en el parell (28), però no el (29). Nogensmenys, hi ha casos en què la relació sí es pot donar en les dues direccions. Es tracta d'*entailment* bidireccional. Vegem-ne un exemple:

30. T/H- *Pedro es el hermano de Patricia*.
    30a. H/T - *Patricia y Pedro son hermanos.*

Tant si la primera frase (30) és el text i la segona (30a) és la hipòtesi com en el cas contrari, es compleix que si el primer és cert, el segon és definitivament cert. Aquestes oracions també són paràfrasis l'una de l'altra. La contradicció, en canvi, no té direcció.

4. Conclusions

En aquest treball s'ha descrit la metodologia per a la creació i anotació de parells de text-hipòtesi per a un corpus d'*entailment* en espanyol. Especialment, s'han definit els vuit criteris que s'han aplicat en la creació de les hipòtesis i s'han donat exemples de cadascun d'aquests criteris. També s'han comentat i exemplificat les dificultats amb què ens hem trobat, ja que es tracta d'una tasca més complexa del que a priori semblava.

El corpus d'*entailment* en espanyol forma part d'un corpus més gran que s'està duent a terme en el Centre de Llenguatge i Computació (CLiC) de la Universitat de Barcelona. Serà el primer corpus amb aquestes característiques per a l'espanyol: fet amb una metodologia que garanteix la complexitat, la riquesa en la variació i amb informació no esbiaixada. Encara està en procés de desenvolupament, ja que volem obtenir un corpus de mida més gran que es pugui implementar en tasques d'aprenentatge automàtic.

Els criteris que s'han donat en aquest treball volen servir de guia per a la creació de corpus d'*entailment* i contradicció fet per humans en altres llengües, per tal de garantir la qualitat i riquesa dels corpus a través de la variació. De cara al futur, seria molt interessant aplicar aquest projecte en una altra llengua, ja sigui en una que tingui únicament un corpus de *Natural Language Inference* (NLI) obtingut aplicant tècniques de traducció, com era el castellà, o en una llengua en què no s'hagi tractat encara la NLI. Un exemple del darrer podria ser el català.

Aquest projecte no només ha servit per crear un corpus, sinó que m'ha permès veure com es desenvolupa en totes les seves fases, des del disseny fins a l'anotació. He treballat amb un grup de professionals que m'han ensenyat com es treballa en un projecte col·laboratiu i he vist la importància de la revisió de criteris i la seva posterior modificació per obtenir el millor resultat possible.

5. Bibliografia

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 632-642. https://doi.org/10.18653/v1/d15-1075

Conneau, A., Rinott, R., Lample, G., Schwenk, H., Stoyanov, V., Williams, A., & Bowman, S. R. (2020). XNLI: Evaluating cross-lingual sentence representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 2475-2485. https://doi.org/10.18653/v1/d18-1269

Crystal, D. (1998). Dictionary of Linguistics and Phonetics. *Language*. https://doi.org/10.2307/417640

Dagan, I., Dolan, B., Magnini, B., & Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, *15*(4). https://doi.org/10.1017/S1351324909990209

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3944 LNAI*, 177-190. https://doi.org/10.1007/11736790_9

De Marneffe, M. C., Rafferty, A. N., & Manning, C. D. (2008). Finding contradictions in text. *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 1039-1047.

Esplà-Gomis, M., Śanchez-Martinez, F., & Forcada, M. L. (2012). UAlacant: Using online machine translation for cross-lingual textual entailment. **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, *2*, 472-476.

Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., & Dolan, B. (2008). The Fourth PASCAL Recognizing Textual Entailment Challenge. *Proceedings of the First Text Analysis Conference*.

Haag, M., Akmajian, A., Demers, R. A., Farmer, A. K., & Harnish, R. M. (1997). Linguistics: An Introduction to Language and Communication. *Language*.

https://doi.org/10.2307/417355

Naranjo, P. (2017). Diccionario de sinónimos y antónimos. Recuperat 5 febrer 2020, de https://www.lenguaje.com/herramientasV2/sinonimos.html

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., & Giampiccolo, D. (2012). Semeval-2012 Task 8: Cross-lingual textual entailment for content synchronization. **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, *2*, 399-407.

Negri, M., & Mehdad, Y. (2010). Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush. *Naacl Hlt 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, (June), 212-216.

Peñas, A., Rodrigo, Á., & Verdejo, F. (2006). SPARTE, a test suite for recognising textual entailment in Spanish. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3878 LNCS*(June 2014), 275-286. https://doi.org/10.1007/11671299_29

Williams, A., Nangia, N., & Bowman, S. R. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. New Orleans, Louisiana.

Zhao, J., Lan, M., & Niu, Z.-Y. (2013). ECNUCS : Recognizing Cross-lingual Textual Entailment Using Multiple Text Similarity and Text Difference Measures. *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, *2: Seventh*(SemEval), 118-123.

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

Tel. +34 934 035 594
fil-coord@ub.edu
www.ub.edu

Declaració d'autoria

Amb aquest escrit declaro que sóc l'autor/autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18, del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 12 de juny de 2020

Signatura: