Trabajo Final de Inteligencia Artificial

Resolución de anáfora de 3ra persona en varios idiomas

Altamirano, Ivana Romina

Marzo de 2011

Índice

Introducción ….....3

Trabajo previo ….....4

Algoritmo de Hobbs.....5

Implementación.....….....7

•Entorno experimental

Idiomas estudiados….....10

▪Corpus del español y del catalán

▪Corpus del italiano

Resultados.....11

Análisis de errores….....12

Conclusiones y trabajo futuro.....20

Herramientas disponibles ….....21

Bibliografía ….....22

Introducción

Los pronombres son palabras o morfemas cuyo referente no es fijo sino que se determina según otras palabras que ya se han nombrado, o incluso pueden no haberse nombrado aún.
A esta propiedad de referirse a otros elementos tanto lingüísticos como extralingüísticos se la denomina deíxis; por ejemplo:

Pedro no es tan listo como él pensaba.
ellos comen mucho.

El pronombre es una palabra cuyo referente depende del contexto lingüístico, al carecer de significado léxico. La diferencia entre las anáforas y los pronombres es que las primeras requieren un antecedente en su dominio sintáctico local, que lo «rija» propiamente de acuerdo con la teoría de rección y ligamiento dentro del marco generativista. Los pronombres por el contrario no requieren la presencia de un antecedente sintáctico, aunque generalmente sí de un antecedente discursivo.
La identificación de correferencias es útil en traducción automática, en respuesta a preguntas en resúmenes automáticos, etc.
Para hacer un sistema de identificación de correferencia que funcione para varios idiomas se deberían estudiar las características particulares de los correferentes a travez de los idiomas.
Algunas cosas a tener en cuenta son: si la lengua distingue o no plurales, (el número) y si tiene género gramatical.
En español los pronombres personales expresan obligatoriamente persona, género, número y caso.
Los pronombres posesivos expresan lo anterior excepto el caso, y el resto de pronombres generalmente sólo expresan género y número. Los pronombres del español pueden ser clasificados en:

- personales (ella, él, ellos, ellas ...)
- tónicos (no-clíticos) (de casos nominativo-vocativo y preposicional) (yo, tú, él, ... mí, ti, sí, ...)
- átonos (clíticos) (de casos acusativo y dativo) (me, te, se; le, lo, la; nos, os; ...)
- reflexivos y recíprocos (me, te, se, nos, os)
- posesivos (mio, tuyo ...)
- demostrativos (eso, esa ...)
- relativos
- interrogativos o exclamativos (enfáticos)
- indefinidos (todos, ninguno)
- numerales

Para realizar este proyecto se estudió el algoritmo de Hobbs (naive), se desarrolló un algoritmo similar adaptado a los corpus brindados por SemEval2011, para estudiar los pronombres personales 3ra persona, en varios idiomas. Se mejoró el algoritmo para que permita resolver referentes elididos (muy comunes en español, pero no en inglés). Se mostrarán resultados en español, catalán e italiano.

Trabajo previo

La correferencia a sido muy estudiada, y hay muchísimas aproximaciones a la resolución muchas con mucha teoría por detras, aquí se nombrarán las principales líneas de investigación.
En 1978 Jerry Hobbs crea 2 algoritmos el Naive que sólo usaba información morfológica y el árbol de parsing sintáctico y otro con información de preferencias semánticas, para resolver correferencias de los pronombres personales 3ra persona él prueba su algoritmos en textos de historias, novelas y noticias (todos en inglés) y reporta un score de 88,3% y 91% para el naive y el mejorado con preferencias semánticas respectivamente.
Luego en 1986 Grosz and Sidner desarrollan la teoría de centering, el texto tiene una estructura lingüística, una estructura intencional, y un estado de foco, las entidades son los centros que unen las oraciones en el discurso. Brennan, Friedman and Pollard(BFP) en el 1987 basado en centering, ranking y filtros. Luego en el 1989 Walker hace una evaluación manual del algoritmo de Hobbs versus el BFP , él concluye que ambos algoritmos se comportaron igual en un dominio de ficción, pero que en noticias de diarios el algoritmo de Hobbs fue mejor que el BFP con un score de 89% contra 79%.
En 1998 Strube estudia S-list (salience list) que describe el estado de la atención en el discurso. El compara BFP con S-list en historias cortas y artículos de diarios (inglés) y obtiene 76% y 85% para BFP y S-list respectivamente. Luego en el 99 Tetreault hacen un algoritmo para resolución incremental de los pronombres (LRC), este algoritmo supero el score del algoritmo de Hobbs, del BFP y del S-list para artículos del New York Times y en textos de ficción.
Para el español en 1999 Palomar, M.; Ferrández, A.; Moreno, L.; hacen un algoritmo en prolog, usan SUG (Slot unification gramars) y full parsing para resolver las correferencias, ellos también aplican su algoritmo al inglés y obtiene un score de 87% para el inglés y 83% para el español, con textos de The Blue Book y en el 2001 Palomar, M.; Ferrández, A.; Moreno, L.; Martínez-Barco, P.; Peral, J.; Muñoz,R.; Saiz-Noeda, M.; hacen un estudio profundo de los pronombres (diferenciando tipos de pronombres) cuando estan o no en frases preposicionales o en frases nominales y definen las restricciones y preferencias según ese estudio, ellos también estudian la no correferencia, el pronombre vacío. Para evaluar sus resultados usaron el Blue Book (manuales técnicos) y en el Lexesp corpus (textos literarios) ellos reportan 76.8% de escore, teniendo el mejor score de 92% para los pronombres reflexivos.
En el 2002 Vincent Ng and Claire Cardie usan machine learning para resolución de correferencias reportan una f-measure de 70,4% y 63,4% respectivamente en el MUC-6 y MUC-7.
Luego en el 2009 Altaf Rahman and Vincent Ng usan modelos no supervizados para la resolución de correferencias prueban su aproximación con el corpus ACE-2005.

Algoritmo de Hobbs

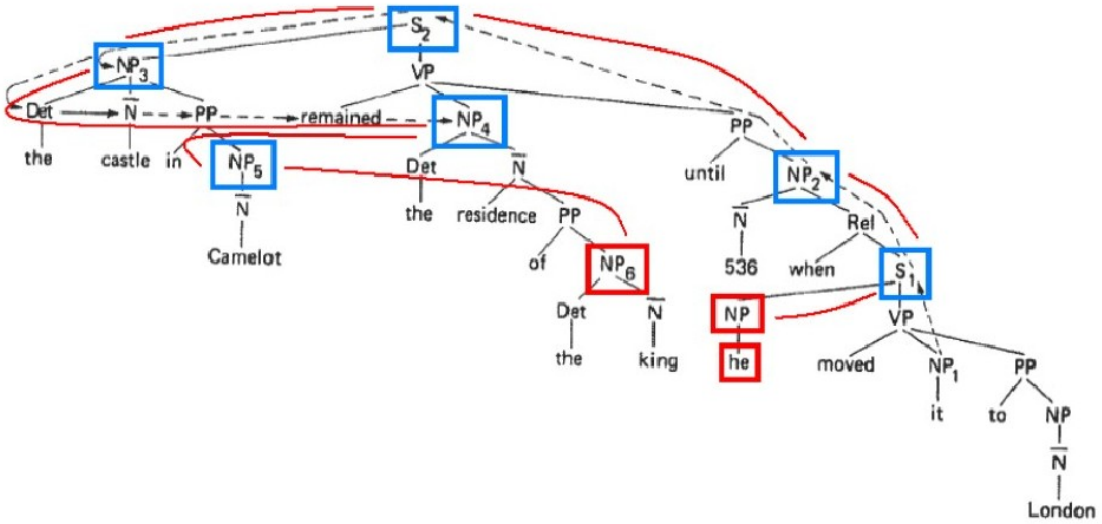
El algoritmo (naive) de Hobbs necesita que tengamos el árbol sintáctico de las oraciones, y además información morfológica, ya que para dar a una entidad como correferente de un pronombre, chequea que ésta sea del mismo género y número de la referencia. Hay otra aproximación de Hobbs, que usa información semántica. El algoritmo de Hobbs esta hecho para resolver referencias pronominales de 3ra persona.

Veamos el algoritmo naive (el que usasolo información morfológica):

1- Empiezo en el nodo NP que inmediatamente domina al pronombre
2- subo en el árbol sintáctico de la oración hasta el primer nodo NP o S encontrado, llamo a este X y al camino para alcanzarlo p
3- Veo las ramas de izq a der en breadth-first. Propone como antecedente cualquier NP el cual tiene un nodo NP o S entre NP y X
4- Si el nodo X es el más grande nodo S en el sentencia, ir de izq a der en breadth-first el árbol de parsing de las sentencias previas en el texto. Cuando un NP es encontrado lo propongo como antecedente (si coincide en género y número), si X no es el nodo más grande S continuo en 5-
5- Desde X vaya arriba en el árbol al primer NP o S encontrado, llame a este X, y al path p.
6- Si X es un NP(y coincide en género y número) y si el path p a X no paso atravez de N' que domina inmediatamente a X, proponga X como antecedente.
7- Atraviece todas las ramas que pertenecen a X a la izq de p, de izq a der y en breadth-first, proponga cualquier NP encontrado (si coincide en género y número) como antecedente.
8- Si X esta en S atraviece todas las ramas de X de la derecha de p, de izq a der en modo breadth- first, pero no vaya atravez de cualquier NP o S encontado.
Proponga cualquier NP (si coincide en género y número) encontrado como antecedente
9- Vaya a 4

Aquí mostramos un ejemplo, que es un poquito mejorado al algoritmo dicho anteriormente ya que tiene preferencias semánticas.
La oración en inglés es: The castle in Camelot remained the residence of the king until 536 when he moved it to London
Tenemos 2 pronombres: it, y he, y a continuación la solución para he, la de it es similar pero elije “the residence” por ser la primera que coincide en género y número y respeta las restricciones preferenciales (ej, grandes construcciones no se pueden mover, lo que descarta the castle por ejemplo)

Comienza en el nodo NP, que esta sobre “he”, el paso 2 lo lleva a S_1 (dice que $X=S_1$, y marca como p, al path para alcanzarlo), el paso 3 busca a la izquierda de p, no encuentra nada, 4 no aplica por no ser S_1 la raíz, 5 hace subir hasta NP_2, ($X=NP_2$ p=path para alcanzarlo), 6 536 es un $\overline{N}$, lo propondría pero lo filtra con preferencias seleccionales, ya que 536 no es algo que se pueda mover. 8 no se cumple, 9 retorna a 4, 4 no se aplica 5 lo lleva a S_2, el paso 6 no se cumple, el paso 7 propone NP_3, pero castle no puede ser la palabra para el pronombre he, por lo tanto lo rechaza por no cumplir con preferencias seleccionales. Y así siguiendo, también rechaza “residence” por la misma razón que castle, y acepta NP_6 “king”.



En estos dos casos vemos como las “preferencias seleccionales” hacen funcionar al algoritmo, el cual sin preferencias seleccionales hubiera dado un referente falso en ambos casos.
Para el pronombre “he” prueba 536, castle, Camelot, residence y finalmente king
Para el pronombre “it” prueba 536, castle, Camelot y finalmente residence

Implementación

Entorno experimental

El entorno experimental es el del SemEval2010, ellos proveían corpus para 6 idiomas, el español, el catalán, el italiano, el inglés, el neerlandes y el alemán, todos los corpus estaban en formato txt, y las anotación variaba según el idioma en algunos casos había anotación manual y automática para varios campos, en otros casos sólo había información automática, los corpus era distintos corpus no eran textos alineados entre los idiomas, tambén diferían en cuanto a la anotación disponible, los corpus tenían cantidad variable de columnas separadas por /t, eran variables por las columnas semánticas que podían variar en cantidad, todos los corpus de training tenían anotación manual de entidades y correferencia. Los corpus de test, no proveían estas últimas anotaciones nombradas. La competencia también provee un conjunto de herramientas de scoring.

La tarea del SemEval:
•Encontrar las expresiones referenciales
•Resolver la correferencia

La tarea realizada en este proyecto: la utilización de dichos corpus de training para resolver la correferencia de los pronombres personales 3ra persona usando el algoritmo de Hobbs modificado para recibir los datos de los corpus del SemEval. También se pretende mejorar el algoritmo para el español y mostrar líneas futuras de investigación.

A modo ilustrativo se muestra una oración en español:

1	EDF	EDF	EDF	n	n	postype=proper	postype=proper	17	17	suj	suj
	(org	_	_	_	_	_	arg0-agt	_	_	_	_	arg0-agt	_
	_	(1											
2	,	,	,	f	f	punct=comma	punct=comma	5	5	f	f
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											
3	que	que	que	p	p	postype=relative|gen=c|num=c	postype=relative|gen=c|num=c	5	5
	suj	suj	_	_	_	_	arg0-agt	_	_	_	_	arg0-agt	_
	_	_	_	(1)									
4	no	no	no	r	r	postype=negative	postype=negative	5	5	mod	mod
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											
5	quiso	querer	querer	v	v	postype=main|num=s|person=3|mood=indicative|tense=past	postype=main|gen=c|num=s|
person=3|mood=indicative|tense=past	1	1	S	S	_	_	querer.a2	querer.a2	_	_	_
	_	_	_	_	_	_	_	_	_	_	_	_	_
6	revelar	revelar	revelar	v	v	postype=main|mood=infinitive	postype=main|gen=c|num=c|mood=infinitive	9
	5	infinitiu	cd	_	_	_	_	_	_	_	_	_	_
	_	_	_	_	_								
7	cuánto	cuánto	cuánto	p	p	postype=interrogative|gen=m|num=s	postype=interrogative|gen=m|num=s	9	5
	cd	cd	_	_	_	_	_	arg1-pat	_	_	_	_	_
	_	_	_	_									
8	_	_	_	p	p	_	_	9	9	suj	suj	_	_
	_	_	_	arg0-agt	_	_	_	_	arg0-agt	_	_	_	(1)
9	pagó	pagar	pagar	v	v	postype=main|num=s|person=3|mood=indicative|tense=past	postype=main|gen=c|num=s|
person=3|mood=indicative|tense=past	5	5	cd	S	_	_	pagar.a32	pagar.a32	arg1-pat	_	_
	_	_	arg1-pat	_	_	_	_	_					
10	por	por	por	s	s	postype=preposition	postype=preposition	9	9	cc	cc
	_	_	_	_	_	argM-adv	_	_	_	_	argM-cau	_	_
	_	_											
11	su	su	su	d	d	postype=possessive|gen=c|num=s|person=3	postype=possessive|gen=c|num=s|person=3
	12	12	spec	spec	_	_	_	_	_	_	_	_	_
	_	_	_	_	_	(1)|(5							
12	participación		participación		participación	n	n	postype=common|gen=f|num=s	postype=common|gen=f|
num=s	10	10	sn	sn	_	_	_	_	_	_	_	_	_
	_	_	_	_	_								
13	mayoritarias.a	mayoritarios.a	mayoritario	a	a	postype=qualificative|gen=f|num=s	postype=qualificative|gen=f|num=s	12	12
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_	_	_									
14	en	en	en	s	s	postype=preposition	postype=preposition	12	12	sp	sp
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											
15	EAA	EAA	EAA	n	n	postype=proper	postype=proper	14	14	sn	sn
	(org)	_	_	_	_	_	_	_	_	_	_	_	_

	_	(2)|5)										
16	,	,	,	f	f	punct=comma		punct=comma		5	9	f	f
	org)	_	_	_	_	_	_	_	_	_	_	_	_
	_	1)											
17	intervendrá	intervenir	intervenir	v	v	postype=main|num=s|person=3|mood=indicative|tense=future					postype=main|gen=c|num=s|		
person=3|mood=indicative|tense=future				0	0	sentence	sentence	_	_	intervenir.a2	intervenir.a2	_	_
	_	_	_	_	_	_	_	_	_				
18	como	como	como	c	c	postype=subordinating		postype=subordinating		17	17	cc	cc
	_	_	_	_	_	_	argM-adv	_	_	_	_	argM-adv	_
	_	_											
19	asistente	asistente	asistente	n	a	postype=common|gen=c|num=s			postype=qualificative|gen=c|num=s			18	18
	sn	sn	_	_	_	_	_	_	_	_	_	_	_
	_	_	_	_									
20	en	en	en	s	s	postype=preposition		postype=preposition		17	17	creg	cc
	_	_	_	_	_	_	arg1-loc	_	_	_	_	arg2-null	_
	_	_											
21	la	el	el	d	d	postype=article|gen=f|num=s			postype=article|gen=f|num=s			22	22
	spec	spec	_	_	_	_	_	_	_	_	_	_	_
	_	_	_	(6									
22	construcción		construcción		construcción		n	n	postype=common|gen=f|num=s			postype=common|gen=f|	
num=s	20	20	sn	sn	_	_	_	_	_	_	_	_	_
	_	_	_	_	_	_							
23	de	de	de	s	s	postype=preposition		postype=preposition		22	22	sp	sp
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											
24	Altamira_2	Altamira_2	Altamira_2	n	n	postype=proper		postype=proper		23	23	sn	sn
	(loc)	_	_	_	_	_	_	_	_	_	_	_	_
	_	(3)|6)											
25	y	y	y	c	c	postype=coordinating		postype=coordinating		17	17	coord	coord
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											
26	,	,	,	f	f	punct=comma		punct=comma		27	27	f	f
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											
27	posteriormente		posteriormente		posteriormente		r	r		_	31	31	cc
	cc	_	_	_	_	_	_	_	argM-tmp	_	_	_	_
	argM-adv	_	_										
28	,	,	,	f	f	punct=comma		punct=comma		27	27	f	f
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											
29	_	_	_	p	p	_	_	31	31	suj	suj	_	_
	_	_	_	_	_	arg0-agt	_	_	_	_	arg0-agt	_	(1)
30	se	él	se	p	p	gen=c|num=c|person=3		gen=c|num=c		31	31	morfema.pronominal	
	morfema.pronominal		_	_	_	_	_	_	_	_	_	_	_
	_	_	_	_									
31	encargará	encargar	encargar	v	v	postype=main|num=s|person=3|mood=indicative|tense=future					postype=main|gen=c|num=s|		
person=3|mood=indicative|tense=future				17	17	S	S	_	_	encargar.a2	encargar.a2	_	_
	_	_	_	_	_	argM-cau	_	_	_				
32	de	de	de	s	s	postype=preposition		postype=preposition		31	31	creg	creg
	_	_	_	_	_	_	_	arg1-null	_	_	_	_	arg2-null
	_	_											
33	explotarla	explotar	explotar+lo	v	v	postype=main|mood=infinitive			postype=main|gen=c|num=c|mood=infinitive				32
	32	S	S	_	_	explotar.a2	explotar.a2	_	_	_	_	_	_
	_	_	_	_	_								
34	como	como	como	c	c	postype=subordinating		postype=subordinating		33	33	cc	cc
	_	_	_	_	_	_	_	_	argM-adv	_	_	_	_
	argM-adv	_											
35	principal	principal	principal	a	a	postype=qualificative|gen=c|num=s			postype=qualificative|gen=c|num=s			36	34
	s.a	sn	_	_	_	_	_	_	_	_	_	_	_
	_	_	_	_									
36	accionista	accionista	accionista	a	a	postype=qualificative|gen=c|num=s			postype=qualificative|gen=c|num=s			34	35
	sn	s.a	_	_	_	_	_	_	_	_	_	_	_
	_	_	_	_									
37	.	.	.	f	f	punct=period		punct=period		17	17	f	f
	_	_	_	_	_	_	_	_	_	_	_	_	_
	_	_											

Ya que ibamoa a implementar el aglgoritmo de Hobbs naive, se filtraron las columnas que iban a ser usadas (se prefirieron las manuales, en el caso de haber manual y automática) y se armó un XML con los datos, esto se realizó solamente por motivos de visualización.
Quedando la misma oración de la siguiente manera:

```
<oracion id_oracion='4' cantidad_pal='37'>
                    <palabra id='1' pal='EDF' p_of_speach='n' morpho='postype=proper' head='17' rel_dep='suj' ent='(org' corref='(1'></palabra>
                    <palabra id='2' pal=',' p_of_speach='f' morpho='punct=comma' head='5' rel_dep='f' ent='_' corref='_'></palabra>
                    <palabra id='3' pal='que' p_of_speach='p' morpho='postype=relative|gen=c|num=c' head='5' rel_dep='suj' ent='_' corref='(1)'></palabra>
                    <palabra id='4' pal='no' p_of_speach='r' morpho='postype=negative' head='5' rel_dep='mod' ent='_' corref='_'></palabra>
                    <palabra id='5' pal='quiso' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=past' head='1' rel_dep='S' ent='_'
corref='_'></palabra>
                    <palabra id='6' pal='revelar' p_of_speach='v' morpho='postype=main|mood=infinitive' head='9' rel_dep='infinitiu' ent='_' corref='_'></palabra>
                    <palabra id='7' pal='cuánto' p_of_speach='p' morpho='postype=interrogative|gen=m|num=s' head='9' rel_dep='cd' ent='_' corref='_'></palabra>
                    <palabra id='8' pal='_' p_of_speach='p' morpho='_' head='9' rel_dep='suj' ent='_' corref='(1)'></palabra>
                    <palabra id='9' pal='pagó' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=past' head='5' rel_dep='cd' ent='_'
corref='_'></palabra>
                    <palabra id='10' pal='por' p_of_speach='s' morpho='postype=preposition' head='9' rel_dep='cc' ent='_' corref='_'></palabra>
                    <palabra id='11' pal='su' p_of_speach='d' morpho='postype=possessive|gen=c|num=s|person=3' head='12' rel_dep='spec' ent='_' corref='(1)|
```

(5'></palabra>
<palabra id='12' pal='participación' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='10' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='13' pal='mayoritaria' p_of_speach='a' morpho='postype=qualificative|gen=f|num=s' head='12' rel_dep='s.a' ent='_' corref='_'></palabra>
<palabra id='14' pal='en' p_of_speach='s' morpho='postype=preposition' head='12' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='15' pal='EAA' p_of_speach='n' morpho='postype=proper' head='14' rel_dep='sn' ent='(org)' corref='(2)|5)'></palabra>
<palabra id='16' pal=',' p_of_speach='f' morpho='punct=comma' head='5' rel_dep='f' ent='org)' corref='1)'></palabra>
<palabra id='17' pal='intervendrá' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=future' head='0' rel_dep='sentence' ent='_' corref='_'></palabra>
<palabra id='18' pal='como' p_of_speach='c' morpho='postype=subordinating' head='17' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='19' pal='asistente' p_of_speach='n' morpho='postype=common|gen=c|num=s' head='18' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='20' pal='en' p_of_speach='s' morpho='postype=preposition' head='17' rel_dep='creg' ent='_' corref='_'></palabra>
<palabra id='21' pal='la' p_of_speach='d' morpho='postype=article|gen=f|num=s' head='22' rel_dep='spec' ent='_' corref='(6'></palabra>
<palabra id='22' pal='construcción' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='20' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='23' pal='de' p_of_speach='s' morpho='postype=preposition' head='22' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='24' pal='Altamira_2' p_of_speach='n' morpho='postype=proper' head='23' rel_dep='sn' ent='(loc)' corref='(3)|6)'></palabra>
<palabra id='25' pal='y' p_of_speach='c' morpho='postype=coordinating' head='17' rel_dep='coord' ent='_' corref='_'></palabra>
<palabra id='26' pal=',' p_of_speach='f' morpho='punct=comma' head='27' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='27' pal='posteriormente' p_of_speach='r' morpho='_' head='31' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='28' pal=',' p_of_speach='f' morpho='punct=comma' head='27' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='29' pal='_' p_of_speach='p' morpho='_' head='31' rel_dep='suj' ent='_' corref='(1)'></palabra>
<palabra id='30' pal='se' p_of_speach='p' morpho='gen=c|num=c|person=3' head='31' rel_dep='morfema.pronominal' ent='_' corref='_'></palabra>
<palabra id='31' pal='encargará' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=future' head='17' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='32' pal='de' p_of_speach='s' morpho='postype=preposition' head='31' rel_dep='creg' ent='_' corref='_'></palabra>
<palabra id='33' pal='explotarla' p_of_speach='v' morpho='postype=main|mood=infinitive' head='32' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='34' pal='como' p_of_speach='c' morpho='postype=subordinating' head='33' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='35' pal='principal' p_of_speach='a' morpho='postype=qualificative|gen=c|num=s' head='36' rel_dep='s.a' ent='_' corref='_'></palabra>
<palabra id='36' pal='accionista' p_of_speach='a' morpho='postype=qualificative|gen=c|num=s' head='34' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='37' pal='.' p_of_speach='f' morpho='punct=period' head='17' rel_dep='f' ent='_' corref='_'></palabra>
</oracion>

Implementamos el algoritmo de Hobbs pero que en vez de tomar el árbol sintáctico toma información de dependencias sintácticas, por la naturaleza de los corpus seleccionados (en los corpus proveen para cada palabra su “head” y su relación de dependencia sintáctica, el head indica de cual palabra depende, y la relación de dependencia sintáctica la función por la cual depende). Las palabras cuyo head es 0, son las raices del árbol, normalmente un verbo. Notar que este árbol es diferente del árbol de parsing pero vamos a recorrer este árbol de manera de recorrerlo de la misma manera que Jerry Hobbs recorre el árbol de parsing, para ello vamos a chequear que la secuencia de sustantivos visitados sea la misma.

Como el algoritmo que se va a hacer permitirá ser unsado en diferentes idiomas definimos REQ que dependerá de la lengua particular estudiada.

REQ=que la palabra estudiada sea un sustantivo, esto se puede dar por varias etiquetas morfosintáctica, y si la lengua estudiada tiene género y/o número para los sustantivos y pronombres, entonces ambos deberán concordar en esta información.

1-Recorro la oración hasta encontrar un pronombre (un cierto nivel en el árbol) sea X
2-Si el head de X fuera 0 (notar que no hay palabra con id=0) estamos en el primer nivel del árbol
    Busco entre los hermanos de X (es decir aquellos cuyos head también sea 0) de izquierda a derecha (es decir según id de menor a mayor)
    Si encuentro que hermano(chequea primero los hermanos) o hijo directo de hermano que cumple REQ
        Devuelvo hermano o hijo que cumple REQ
        FIN
    si no
        busco en hijos de hijos de izquierda a derecha
sino
    Subo en el árbol hasta el padre de X y lo llamo A (en el corpus esto esta dado por el head de la palabra, me da el id de su padre)
    Si el head de A es =0
        voy a 2
    sino
        Subo en el árbol hasta el padre de A y lo llamo B
4-Si B cumple REQ

devuelvo B
FIN
sino
Busco entre los hijos de B (solo primer nivel) de izquierda a derecha (id's de las palabras que su head sea igual al id de B)
Si encuentro hijo que cumple REQ
Devuelvo ese hijo
FIN
si no
busco el padre de B y lo llamo X
sigo en 2

Idiomas estudiados

Vamos a estudiar los idiomas de la competencia del Semeval ya que ellos proveen corpus etiquetado con entidades nombradas y correferencias entre otras, esto permitirá que se pueda evaluar la salida de manera automática para los 3 idiomas: catalán, italiano y español

Corpus del español y del catalán – características

Los palabras elididas estan marcadas, con "_" en la posición de la palabra elidida.
Ej: " que no quiso revelar cuánto _ pago por ..."

Estos dos corpus provienen del AnCora (textos de diarios, noticias) y tiene las siguientes características:
Anotación: morfológica (PoS and lemmas), sintactica (constituyentes y funciones), y semántica (argumentos estructuras, roles temáticos, clases semánticas de verbos, Nes (entidades nombradas), sentidos nominales WordNet). Estos corpus proveían algunas características con dos atributos, el manual y el automático. En este proyecto se uso el manual, en todos los casos que había.
Más la anotación de expresiones referenciales, y correferencias provista por el SemEval que era anotación manual.

Identificación de pronombres en catalán y español

Etiqueta part-of-speach "p" indica que es un pronombre, y los sustantivos estan indicados con "n" tienen género y número, dados por f, m y s, p para femenino, masculino, singular, plural respectivamente.

Corpus del italiano - características

Proviene de la wikipedia italiana, parseados y tagueados automáticamente. La anotación de PoS fue la dada por el TextPro y la información de dependencias sintácticas por el MALT Parser
Identificación de pronombres en italiano

El tagset de la etiqueta part-of-speach esta disponible en: http://textpro.fbk.eu/tagset_it.html
de ahi vimos que PS, PP, y PN son pronombres singular, plural y neutro respectivamente.
Identificamos los sustantivos plurales SP, singulares SS, neutros para números SN y SPN propios.

Resultados

Análisis de la aplicación del algoritmo al español y analizando el pronombre “él”

1- Al aplicar el algoritmo de Hobbs, fallaba muchas veces porque “él” refería a un nombre propio, y el nombre propio no tenia ni género ni número, por lo tanto, fallaba.- Arreglado con eurísticas de selección.
Aguegue que pudiera dar verdadero cuando era un nombre propio, pero esto incluyo otros problemas, como ser elegir un país... (ya que es un nombre propio) por lo tanto agregue que no contengan ent='(loc)'
Y asi incluyo un nuevo problema por ejemplo:
“Esta es la provocativa teoría de dos arqueólogos norteamericanos, en abierta contradicción con la generalmente aceptada de que América fue poblada hace 13.500 años por nómadas procedentes de Asia que entraron en ella por Alaska.”
En esa oración “América” una localidad es realmente el referente.

2- Fallaba en encontrar los elididos porque no tienen género ni número proveen esta anotación morpho='_'
Agregue que se permitan los elididos.
Al agregar esto lo que hago es resolver un problema e incluir otro... si la persona referente esta más a la izquierda, claramente no la va a buscar porque se conformará con el primer “_” elidido que encuentre. (lo resolví dejando correr el algoritmo luego de haber encontrado y prefiriendo otras palabras antes que los “_” elididos)
Ej: “_ Estaba advertida y para ella _ no ha sido una sorpresa.”

3- Si un pronombre apunta a un verbo que esta luego de otro sustantivo con las mismas características, falla.

4- Ocurren fallas con elementos lingüísticos tales como “como”, “anadió”, la falla ocurre porque una digamos “aclaración” confunde al algoritmo.

5- Hay errores en el corpus Ej: “La manifestación de hoy ha estado precedida en las últimas semanas por marchas similares, aunque ninguna de ella con la convocatoria...” oración 2313 aparece el pronombre ella, el cual refiere a marchas, pero no concuerda en número.

6- Puede fallar cuando no tiene correferente o el correferente se encuentra a la derecha del pronombre “En_nombre_de ella se consiguió poner_en_marcha un sistema multipartidista y parlamentario, pero en los últimos años este sistema se ha resquebrajado por el auge del islamismo interno y por la presión de una minoría kurda que acabó recurriendo al terrorismo.”
Realicé una modificación que fue dejar seguir corriendo el algoritmo luego de que encontrara algún referente, porque en español al haber muchos sujetos elididos, no encuentra el referente, notar que podría encontrar el referente y luego un elidido, por lo tanto ahora necesito eurísticas de selección (pues ahora tenemos varios candidatos), ellas son:
a). Suponiendo que el pronombre es “él” preferir un nombre propio en vez de un “_” elidido, porque es probable que el elidido siga en cadena del nombre propio. Esto puede traer problemas si hay nombre propio entre el pronombre y el referente. Notar que el referente tambien puede ser un sustantivo común.
b). Si el pronombre fuera “eso” preferiría un sustantivo común en vez de uno propio.

Ahora voy a presentar los resultados de la ejecusión del algoritmo buscando pronombres personales

3ra persona y sus referentes, comparados con la correferencia provista por el corpus para 3 idiomas.
En caso de éxito, es decir el referente encontrado esta en marcado manualmente en el corpus para un pronombre particular, suma 1, si no encuentro, no cuento nada, es decir no le doy acierto, aunque el correferente manual sea nulo (es decir no tiene correferente). Al lado sumando los aciertos de no haber encontrado correferente.

Idioma ESPAÑOL	*Exitos sin/con aciertos de referencia vacia*	*Cantidad de pronombres encontrados*	*Precisión*
él	85, 87	117	0,72 – 0,74
ellos	55, 62	100	0,55 – 0,62
ella	24, 26	44	0,54 – 0,59
ellas	20, 20	29	0,69 – 0,69
TOTALES	184, 195	290	**0,63 – 0,67**

Para el pronombre "él" conseguimos mucha mejor precisión y esto se debe a que este refiere muchas veces a nombres propios. No sucede lo mismo con ella, ellos ellas. En el caso de los plurales muchas fallas pueden venir de enumeraciones de cosas que son todas singulares.
Idioma CATALÁN

Idioma CATALÁN	*Exitos sin/con aciertos de referencia vacia*	*Cantidad de pronombres encontrados*	*Precisión*
ell (masc. singular)	38, 41	69	0,55 – 0,59

Idioma ITALIANO	*Exitos sin/con aciertos de referencia vacia*	*Cantidad de pronombres encontrados*	*Precisión*
lei	2, 5	6	0,33 – 0,83
lui	13, 19	25	0,52 – 0,76
loro	2, 6	20	0,1 – 0,3
TOTALES	17, 30	51	**0,33 – 0,58**

En el caso del italiano, solo teníamos etiquetas part-of-spech automáticas de gruesa granularidad, el corpus era muy pequeño, no se realizó evaluación manual de los resultados.
Problemas encontrados en general independiente de lengua, aunque una palabra no tenga referente, el algoritmo tiende a devolver siempre un referente erroneo.

Análisis de Errores

Para el pronombre "él"

Oración 335
(El Gobierno)$_{2}$, apuntó (Pimentel)$_{22}$, se ha enterado de (la operación)$_{10}$, pero _ no ha participado en (la gestión del (pacto entre ((el BBVA)$_{23}$ y Telefónica)$_{11}$)$_{10}$)$_{67}$ y (lo que _ va a hacer ahora)$_{68}$ es estudiar si
(**él** mismo)$_{4}$ vulnera o no (algún tipo de reglas de (la competencia)$_{25}$)$_{69}$.

El algoritmo aca prueba Gobierno que esta como nombre propio, porque esta con mayúscula, y los nombres propios estan permitidos para ser correferente de "él", pero luego lo rechaza por no estar anotado como correferente (en la comparación para ver si sumo o no 1).
Este es un error de anotación del corpus, debería estar anotado así para que anduviera de esta manera

el algoritmo:

(El Gobierno)$_{2}$, apuntó (Pimentel)$_{22}$, se ha enterado de (la operación)$_{10}$, pero _ no ha participado en (la gestión del (pacto entre ((el BBVA)$_{23}$ y Telefónica)$_{11}$)$_{10}$)$_{67}$ y (lo que _ va a hacer ahora)$_{68}$ es estudiar si (**él(2)** mismo)$_{4}$ vulnera o no (algún tipo de reglas de (la competencia)$_{25}$)$_{69}$.

Oración 338
"El señor González - añadió _ - es consciente de que _ sabemos todos los detalles, incluidos los que él mismo ha hablado sobre sí mismo" EFE.

La anotación de 'él' es:

pal='él' p_of_speach='p' morpho='postype=personal|gen=m|num=s|person=3' head='25' rel_dep='suj' ent='_' corref='(4'

En este caso falla por la anotación de “corref”, porque en realidad se esta anotando que “él mismo” y no “él”, entonces encuentra el referente pero como corref no es “(4)”, no lo da como válido.

Oración 1094
Los compañeros se solidarizaron con él, _ se negaron a entrenar el pasado jueves y _ culparon de lo ocurrido a los directivos del cuadro universitario, problemas que con toda seguridad influyeron en el magro resultado de hoy, domingo.

En este caso no hay correferente, pero el corpus tiene anotado 'él' con corref='(21)' por si va a referir en otro lugar. Por lo tanto, el sistema da que no encuentra correferente, lo cual es cierto, pero da false por que ve corref='(21)'.

Oración 1187
"_ No lo será - añadió _ -, pero si él o Nicolás_Redondo están como lehendakari, se supone que no habrá tiros y _ no matarán a nadie".

En esta oración


```
<oracion id_oracion='1187' cantidad_pal='33'>
<palabra id='1' pal='"' p_of_speach='f' morpho='punct=quotation' head='5' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='2' pal='_' p_of_speach='p' morpho='_' head='5' rel_dep='suj' ent='_' corref='(16)'></palabra>
<palabra id='3' pal='No' p_of_speach='r' morpho='postype=negative' head='5' rel_dep='mod' ent='_' corref='_'></palabra>
<palabra id='4' pal='lo' p_of_speach='p' morpho='postype=personal|gen=c|num=c|person=3|case=accusative' head='5' rel_dep='atr' ent='_' corref='_'></palabra>
<palabra id='5' pal='será' p_of_speach='v' morpho='postype=semiauxiliary|num=s|person=3|mood=indicative|tense=future' head='0' rel_dep='sentence' ent='_' corref='_'></palabra>
<palabra id='6' pal='-' p_of_speach='f' morpho='punct=hyphen' head='7' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='7' pal='añadió' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=past' head='5' rel_dep='inc' ent='_' corref='_'></palabra>
<palabra id='8' pal='_' p_of_speach='p' morpho='_' head='7' rel_dep='suj' ent='_' corref='(1)'></palabra>
<palabra id='9' pal='-' p_of_speach='f' morpho='punct=hyphen' head='7' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='10' pal=',' p_of_speach='f' morpho='punct=comma' head='5' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='11' pal='pero' p_of_speach='c' morpho='postype=coordinating' head='5' rel_dep='coord' ent='_' corref='_'></palabra>
<palabra id='12' pal='si' p_of_speach='c' morpho='postype=subordinating' head='16' rel_dep='conj' ent='_' corref='_'></palabra>
<palabra id='13' pal='él' p_of_speach='p' morpho='postype=personal|gen=m|num=s|person=3' head='16' rel_dep='suj' ent='_' corref='(16)|(66'></palabra>
<palabra id='14' pal='o' p_of_speach='c' morpho='postype=coordinating' head='13' rel_dep='coord' ent='_' corref='_'></palabra>
<palabra id='15' pal='Nicolás_Redondo' p_of_speach='n' morpho='postype=proper' head='13' rel_dep='grup.nom' ent='_' corref='(65)|(66)'></palabra>
<palabra id='16' pal='están' p_of_speach='v' morpho='postype=main|num=p|person=3|mood=indicative|tense=present' head='21' rel_dep='ao' ent='_' corref='_'></palabra>
<palabra id='17' pal='como' p_of_speach='c' morpho='postype=subordinating' head='16' rel_dep='atr' ent='_' corref='_'></palabra>
<palabra id='18' pal='lehendakari' p_of_speach='n' morpho='postype=common|gen=c|num=s' head='17' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='19' pal=',' p_of_speach='f' morpho='punct=comma' head='16' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='20' pal='se' p_of_speach='p' morpho='gen=c|num=c' head='21' rel_dep='pass' ent='_' corref='_'></palabra>
<palabra id='21' pal='supone' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=present' head='5' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='22' pal='que' p_of_speach='c' morpho='postype=subordinating' head='24' rel_dep='conj' ent='_' corref='_'></palabra>
```

<palabra id='23' pal='no' p_of_speach='r' morpho='postype=negative' head='24' rel_dep='mod' ent='_' corref='_'></palabra>
<palabra id='24' pal='habrá' p_of_speach='v' morpho='postype=auxiliary|num=s|person=3|mood=indicative|tense=future' head='21' rel_dep='suj' ent='_' corref='_'></palabra>
<palabra id='25' pal='tiros' p_of_speach='n' morpho='postype=common|gen=m|num=p' head='24' rel_dep='cd' ent='_' corref='_'></palabra>
<palabra id='26' pal='y' p_of_speach='c' morpho='postype=coordinating' head='24' rel_dep='coord' ent='_' corref='_'></palabra>
<palabra id='27' pal='_' p_of_speach='p' morpho='_' head='29' rel_dep='suj' ent='_' corref='_'></palabra>
<palabra id='28' pal='no' p_of_speach='r' morpho='postype=negative' head='29' rel_dep='mod' ent='_' corref='_'></palabra>
<palabra id='29' pal='matarán' p_of_speach='v' morpho='postype=main|num=p|person=3|mood=indicative|tense=future' head='24' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='30' pal='a' p_of_speach='s' morpho='postype=preposition' head='29' rel_dep='cd' ent='_' corref='_'></palabra>
<palabra id='31' pal='nadie' p_of_speach='p' morpho='postype=indefinite|gen=c|num=s' head='30' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='32' pal='"' p_of_speach='f' morpho='punct=quotation' head='5' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='33' pal='.' p_of_speach='f' morpho='punct=period' head='5' rel_dep='f' ent='_' corref='_'></palabra>
</oracion>

sucede que el head de 'él' me lleva a 16, luego a 21, luego a 5, 5 me lleva a 0 (con lo cual ve la primer palabra, porque estamos en la raíz) y esta me lleva a 5 nuevamente, que nuevamente me lleva a 0, como ya lo había recorrido, no me devuelve ningún correferente, pero tiene como correferente un elidido, que es la palabra con id='2' según la anotación del corpus.
pal='él' p_of_speach='p' morpho='postype=personal|gen=m|num=s|person=3' head='16' rel_dep='suj' ent='_' corref='(16)|(66'

Oracion 2015
La federación alemana, de la que él es vicepresidente, está buscando ahora un sucesor a Erich_Ribbeck, quien dimitió como seleccionador tras la humillante eliminación de la Eurocopa en la ronda preliminar y escaso juego ofrecido por su conjunto.

Estoy buscando correferente para la palabra: pal='él' p_of_speach='p' morpho='postype=personal|gen=m|num=s|person=3' head='9' rel_dep='suj' ent='_' corref='(1)'

El algoritmo no encuentra referente, pero aparece corref='(1)' que en realidad no es que tenga un correferente, sino que es la primer ocurrencia, y lo anota para correferencia futura.

Oración 2028
"No hay nada que _ pueda hacer por él menos apoyarlo y eso lo que _ pienso hacer", añadió _.
pal='él' p_of_speach='p' morpho='postype=personal|gen=m|num=s|person=3' head='9' rel_dep='sn' ent='_' corref='(2)'

Mismo problema anterior, el algoritmo no encuentra referente, y esto es cierto, porque no tiene referente, pero por la primer ocurrencia aparece etiquetado con corref='(2)' para futuras correferencias y hace que de false.
Oración 2559
Según él, después del trabajo hecho, el Partido_Popular tiene crédito, y "los otros, los que son incapaces de hacerlo, lo máximo que pueden hablar es del reparto".

<oracion id_oracion='2559' cantidad_pal='35'>
<palabra id='1' pal='Según' p_of_speach='s' morpho='postype=preposition' head='11' rel_dep='ao' ent='_' corref='_'></palabra>
<palabra id='2' pal='él' p_of_speach='p' morpho='postype=personal|gen=m|num=s|person=3' head='1' rel_dep='sn' ent='_' corref='(1)'></palabra>
<palabra id='3' pal=',' p_of_speach='f' morpho='punct=comma' head='4' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='4' pal='después' p_of_speach='r' morpho='_' head='11' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='5' pal='del' p_of_speach='s' morpho='postype=preposition|gen=m|num=s|contracted=yes' head='4' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='6' pal='trabajo' p_of_speach='n' morpho='postype=common|gen=m|num=s' head='5' rel_dep='sn' ent='_' corref='(12'></palabra>
<palabra id='7' pal='hecho' p_of_speach='a' morpho='postype=qualificative|gen=m|num=s|posfunction=participle' head='6' rel_dep='S' ent='_' corref='12)'></palabra>
<palabra id='8' pal=',' p_of_speach='f' morpho='punct=comma' head='4' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='9' pal='el' p_of_speach='d' morpho='postype=article|gen=m|num=s' head='10' rel_dep='spec' ent='(org' corref='(3'></palabra>
<palabra id='10' pal='Partido_Popular' p_of_speach='n' morpho='postype=proper' head='11' rel_dep='suj' ent='org)' corref='3)'></palabra>
<palabra id='11' pal='tiene' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=present' head='0'

rel_dep='sentence' ent='_' corref='_'></palabra>
<palabra id='12' pal='crédito' p_of_speach='n' morpho='postype=common|gen=m|num=s' head='11' rel_dep='cd' ent='_' corref='_'></palabra>
<palabra id='13' pal=',' p_of_speach='f' morpho='punct=comma' head='11' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='14' pal='y' p_of_speach='c' morpho='postype=coordinating' head='11' rel_dep='coord' ent='_' corref='_'></palabra>
<palabra id='15' pal='"' p_of_speach='f' morpho='punct=quotation' head='31' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='16' pal='los' p_of_speach='d' morpho='postype=article|gen=m|num=p' head='17' rel_dep='spec' ent='_' corref='(11'></palabra>
<palabra id='17' pal='otros' p_of_speach='p' morpho='postype=indefinite|gen=m|num=p' head='30' rel_dep='suj' ent='_' corref='_'></palabra>
<palabra id='18' pal=',' p_of_speach='f' morpho='punct=comma' head='21' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='19' pal='los' p_of_speach='d' morpho='postype=article|gen=m|num=p' head='21' rel_dep='spec' ent='_' corref='_'></palabra>
<palabra id='20' pal='que' p_of_speach='p' morpho='postype=relative|gen=c|num=c' head='21' rel_dep='suj' ent='_' corref='_'></palabra>
<palabra id='21' pal='son' p_of_speach='v' morpho='postype=semiauxiliary|num=p|person=3|mood=indicative|tense=present' head='17' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='22' pal='incapaces' p_of_speach='a' morpho='postype=qualificative|gen=c|num=p' head='21' rel_dep='atr' ent='_' corref='_'></palabra>
<palabra id='23' pal='de' p_of_speach='s' morpho='postype=preposition' head='22' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='24' pal='hacerlo' p_of_speach='v' morpho='postype=main|mood=infinitive' head='23' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='25' pal=',' p_of_speach='f' morpho='punct=comma' head='21' rel_dep='f' ent='_' corref='11)'></palabra>
<palabra id='26' pal='lo' p_of_speach='d' morpho='postype=article|gen=c|num=s' head='30' rel_dep='spec' ent='_' corref='(18'></palabra>
<palabra id='27' pal='máximo' p_of_speach='a' morpho='postype=qualificative|gen=m|num=s' head='30' rel_dep='s.a' ent='_' corref='_'></palabra>
<palabra id='28' pal='que' p_of_speach='p' morpho='postype=relative|gen=c|num=c' head='30' rel_dep='cd' ent='_' corref='(18)'></palabra>
<palabra id='29' pal='pueden' p_of_speach='v' morpho='postype=main|num=p|person=3|mood=indicative|tense=present' head='30' rel_dep='v' ent='_' corref='_'></palabra>
<palabra id='30' pal='hablar' p_of_speach='v' morpho='postype=main|mood=infinitive' head='31' rel_dep='suj' ent='_' corref='18)'></palabra>
<palabra id='31' pal='es' p_of_speach='v' morpho='postype=semiauxiliary|num=s|person=3|mood=indicative|tense=present' head='11' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='32' pal='del' p_of_speach='s' morpho='postype=preposition|gen=m|num=s|contracted=yes' head='31' rel_dep='atr' ent='_' corref='_'></palabra>
<palabra id='33' pal='reparto' p_of_speach='n' morpho='postype=common|gen=m|num=s' head='32' rel_dep='sn' ent='_' corref='(30)'></palabra>
<palabra id='34' pal='"' p_of_speach='f' morpho='punct=quotation' head='31' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='35' pal='.' p_of_speach='f' morpho='punct=period' head='11' rel_dep='f' ent='_' corref='_'></palabra>
</oracion>

En este caso el algoritmo falla porque al buscar en el árbol los hijos de 11 prueba Partido_Popular que apace como nombre propio y lo acepta. Pero en realidad en este caso no hay referente.

Oración 3375

El posible centro que Microsoft estudia desarrollar en Barcelona sería para telefonía móvil de tercera generación UMTS y en él podrían colaborar Telefónica y la propia UPC.

En esta oración el algoritmo falla porque da ve Telefónica como nombre propio antes que centro el cual esta en el referente.

Oracion 3499

"_ Puedo asegurar que él (Segura) mintió, por_lo_menos a mí", dijo Vázquez_Raña al responder a la decisión del atleta de someterse al detector de mentiras en una cadena de televisión por cable para aclarar que _ ignoraba su descalificación, que ha generado una gran polémica en México.

En este caso el algoritmo no encuentra referente y la anotación no le ayuda, Segura aparece como un nombre propio, pero no aparece anotado como que él refiera a Segura

<oracion id_oracion='3499' cantidad_pal='55'>
<palabra id='1' pal='"' p_of_speach='f' morpho='punct=quotation' head='4' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='2' pal='_' p_of_speach='p' morpho='_' head='4' rel_dep='suj' ent='_' corref='(1)'></palabra>
<palabra id='3' pal='Puedo' p_of_speach='v' morpho='postype=main|num=s|person=1|mood=indicative|tense=present' head='4' rel_dep='v' ent='_' corref='_'></palabra>
<palabra id='4' pal='asegurar' p_of_speach='v' morpho='postype=main|mood=infinitive' head='17' rel_dep='cd' ent='_' corref='_'></palabra>
<palabra id='5' pal='que' p_of_speach='c' morpho='postype=subordinating' head='10' rel_dep='conj' ent='_' corref='_'></palabra>
<palabra id='6' pal='él' p_of_speach='p' morpho='postype=personal|gen=m|num=s|person=3' head='10' rel_dep='suj' ent='_' corref='(3'></palabra>
<palabra id='7' pal='(' p_of_speach='f' morpho='punct=bracket|punctenclose=open' head='8' rel_dep='f' ent='(person' corref='(25'></palabra>
<palabra id='8' pal='Segura' p_of_speach='n' morpho='postype=proper' head='6' rel_dep='sn' ent='_' corref='_'></palabra>

<palabra id='9' pal=')' p_of_speach='f' morpho='punct=bracket|punctenclose=close' head='8' rel_dep='f' ent='person)' corref='25)|3)'></palabra>
<palabra id='10' pal='mintió' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=past' head='4' rel_dep='cd' ent='_' corref='_'></palabra>
<palabra id='11' pal=',' p_of_speach='f' morpho='punct=comma' head='13' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='12' pal='por_lo_menos' p_of_speach='r' morpho='_' head='13' rel_dep='sadv' ent='_' corref='_'></palabra>
<palabra id='13' pal='a' p_of_speach='s' morpho='postype=preposition' head='10' rel_dep='ci' ent='_' corref='_'></palabra>
<palabra id='14' pal='mí' p_of_speach='p' morpho='postype=personal|gen=c|num=s|person=1|case=oblique' head='13' rel_dep='sn' ent='_' corref='(1)'></palabra>
<palabra id='15' pal='"' p_of_speach='f' morpho='punct=quotation' head='4' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='16' pal=',' p_of_speach='f' morpho='punct=comma' head='4' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='17' pal='dijo' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=past' head='0' rel_dep='sentence' ent='_' corref='_'></palabra>
<palabra id='18' pal='Vázquez_Raña' p_of_speach='n' morpho='postype=proper' head='17' rel_dep='suj' ent='(person)' corref='(1)'></palabra>
<palabra id='19' pal='al' p_of_speach='s' morpho='postype=preposition|gen=m|num=s|contracted=yes' head='17' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='20' pal='responder' p_of_speach='v' morpho='postype=main|mood=infinitive' head='19' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='21' pal='a' p_of_speach='s' morpho='postype=preposition' head='20' rel_dep='creg' ent='_' corref='_'></palabra>
<palabra id='22' pal='la' p_of_speach='d' morpho='postype=article|gen=f|num=s' head='23' rel_dep='spec' ent='_' corref='(28'></palabra>
<palabra id='23' pal='decisión' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='21' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='24' pal='del' p_of_speach='s' morpho='postype=preposition|gen=m|num=s|contracted=yes' head='23' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='25' pal='atleta' p_of_speach='n' morpho='postype=common|gen=c|num=s' head='24' rel_dep='sn' ent='_' corref='(3)'></palabra>
<palabra id='26' pal='de' p_of_speach='s' morpho='postype=preposition' head='23' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='27' pal='someterse' p_of_speach='v' morpho='postype=main|mood=infinitive' head='26' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='28' pal='al' p_of_speach='s' morpho='postype=preposition|gen=m|num=s|contracted=yes' head='27' rel_dep='creg' ent='_' corref='_'></palabra>
<palabra id='29' pal='detector' p_of_speach='n' morpho='postype=common|gen=m|num=s' head='28' rel_dep='sn' ent='_' corref='(4'></palabra>
<palabra id='30' pal='de' p_of_speach='s' morpho='postype=preposition' head='29' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='31' pal='mentiras' p_of_speach='n' morpho='postype=common|gen=f|num=p' head='30' rel_dep='sn' ent='_' corref='4)'></palabra>
<palabra id='32' pal='en' p_of_speach='s' morpho='postype=preposition' head='27' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='33' pal='una' p_of_speach='d' morpho='postype=indefinite|gen=f|num=s' head='34' rel_dep='spec' ent='_' corref='(26'></palabra>
<palabra id='34' pal='cadena' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='32' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='35' pal='de' p_of_speach='s' morpho='postype=preposition' head='34' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='36' pal='televisión' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='35' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='37' pal='por' p_of_speach='s' morpho='postype=preposition' head='34' rel_dep='sp' ent='_' corref='_'></palabra>
<palabra id='38' pal='cable' p_of_speach='n' morpho='postype=common|gen=m|num=s' head='37' rel_dep='sn' ent='_' corref='26)'></palabra>
<palabra id='39' pal='para' p_of_speach='s' morpho='postype=preposition' head='27' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='40' pal='aclarar' p_of_speach='v' morpho='postype=main|mood=infinitive' head='43' rel_dep='infinitiu' ent='_' corref='_'></palabra>
<palabra id='41' pal='que' p_of_speach='c' morpho='postype=subordinating' head='43' rel_dep='conj' ent='_' corref='_'></palabra>
<palabra id='42' pal='_' p_of_speach='p' morpho='_' head='43' rel_dep='suj' ent='_' corref='(3)'></palabra>
<palabra id='43' pal='ignoraba' p_of_speach='v' morpho='postype=main|num=s|person=3|mood=indicative|tense=imperfect' head='39' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='44' pal='su' p_of_speach='d' morpho='postype=possessive|gen=c|num=s|person=3' head='45' rel_dep='spec' ent='_' corref='(3)|(5'></palabra>
<palabra id='45' pal='descalificación' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='43' rel_dep='cd' ent='_' corref='_'></palabra>
<palabra id='46' pal=',' p_of_speach='f' morpho='punct=comma' head='49' rel_dep='f' ent='_' corref='_'></palabra>
<palabra id='47' pal='que' p_of_speach='p' morpho='postype=relative|gen=c|num=c' head='49' rel_dep='suj' ent='_' corref='(5)'></palabra>
<palabra id='48' pal='ha' p_of_speach='v' morpho='postype=auxiliary|num=s|person=3|mood=indicative|tense=present' head='49' rel_dep='v' ent='_' corref='_'></palabra>
<palabra id='49' pal='generado' p_of_speach='v' morpho='postype=main|gen=m|num=s|mood=participle' head='45' rel_dep='S' ent='_' corref='_'></palabra>
<palabra id='50' pal='una' p_of_speach='d' morpho='postype=indefinite|gen=f|num=s' head='52' rel_dep='spec' ent='_' corref='(27'></palabra>
<palabra id='51' pal='gran' p_of_speach='a' morpho='postype=qualificative|gen=c|num=s' head='52' rel_dep='s.a' ent='_' corref='_'></palabra>
<palabra id='52' pal='polémica' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='49' rel_dep='cd' ent='_' corref='27)'></palabra>
<palabra id='53' pal='en' p_of_speach='s' morpho='postype=preposition' head='49' rel_dep='cc' ent='_' corref='_'></palabra>
<palabra id='54' pal='México' p_of_speach='n' morpho='postype=proper' head='53' rel_dep='sn' ent='(loc)' corref='(8)|5)|28)'></palabra>
<palabra id='55' pal='.' p_of_speach='f' morpho='punct=period' head='17' rel_dep='f' ent='_' corref='_'></palabra>
</oracion>

Para el pronombre “ella”

Oración 2313

“La manifestación de hoy ha estado precedida en las últimas semanas por marchas similares, aunque ninguna de ella con la convocatoria del cardenal Sin y la presencia, entre otros, de Aquino, el también ex presidente Fidel_Ramos y la vicepresidenta, Gloria_Macapagal-Arroyo, que sustituiría a Estrada si éste deja el poder.”

En este caso el referente es “marchas similares”, pero el algoritmo no la encuentra porque no coincide en género y número.

Oración 4061

“En_nombre_de ella se consiguió poner_en_marcha un sistema multipartidista y parlamentario, pero

en los últimos años este sistema se ha resquebrajado por el auge del islamismo interno y por la presión de una minoría kurda que acabó recurriendo al terrorismo. ”

El algoritmo no encuentra referente, y en realidad no tiene, pero como ocurre anotado con corref='(12)' da false.

Oración 4925
“A ella pertenece el festival internacional de cine documental y cortometraje de Bilbao, donde el filme de Rebollo obtuvo el gran premio del cine español y el de Elizalde, el gran premio del festival de Bilbao y del cine vasco.”

Igual que en el caso anterior, la oración no tiene referente, el algoritmo no encuentra referente, pero ella esta anotado con corref='(9)'

Oración 4989
“_ Trabajo con ella desde 1983.”

Oración 5063
“Fue él quien dio el primer paso y ella dijo sí.”

Oración 5082
“Cuando _ la miro _ no pienso nunca si _ estoy guapa o fea, _ sólo la miro y ella me hace la mitad del trabajo. ”

Oración 5192
“Excepto ella y Romain_Duris, el resto del reparto son los habitantes de ese pueblo rumano. ”

Oracion 5334
“Junto_a ella, estarían los periodistas Jaime_Peñafiel y Antonio_Pérez_Henares.”

Oración 5993
“Es a ella a quien _ he dedicado mi libro.”

Oracion 6237
“"Por ello, _ no la retiramos, pero _ tampoco insistimos en ella", añadió _.”

Para el pronombre “ellos”

Oracion 7
La alcaldesa de Málaga y cabeza de lista del PP al Congreso por esta provincia, Celia_Villalobos, pidió hoy a los militantes de esta formación que _ sepan "administrar la victoria", porque "no vale la revancha, el insulto o el ataque, eso es para ellos".

Estoy buscando correferente para la palabra: pal='ellos' p_of_speach='p' morpho='postype=personal|gen=m|num=p|person=3' head='51' rel_dep='sn' ent='_' corref='(26)'

Aca no encuentro el correferente, porque no esta en el texto, pero lo estoy dando como false, porque al tener esto corref='(26)', y no encontrar ninguno la respuesta difiere. En este caso la corref se refiere a que esta palabra va a ser el referente 26 en el futuro que sea nombrado.

Oracion 197
“Los doce condenados por el asalto al cuartel de La_Tablada, todos ellos miembros del

ultraizquierdista Movimiento_Todos_por_la_Patria (MTP), están hospitalizados a_causa_de una huelga de hambre que _ mantienen desde hace casi 100 días para lograr que se les reconozca el derecho a apelar sus sentencias.”

<palabra id='11' pal=',' p_of_speach='f' morpho='punct=comma' head='13' rel_dep='f' ent='_' corref='(6'></palabra>
<palabra id='12' pal='todos' p_of_speach='d' morpho='postype=indefinite|gen=m|num=p' head='13' rel_dep='spec' ent='_' corref='_'></palabra>
<palabra id='13' pal='ellos' p_of_speach='p' morpho='postype=personal|gen=m|num=p|person=3' head='3' rel_dep='sn' ent='_' corref='6)'></palabra>

Aca el problema es que no encuentra el (en corref, entonces el algoritmo falla, en realidad es que ellos es parte de la referencia.

Oración 287
“El equipo ecuatoriano ya tuvo la oportunidad de enfrentarse a otro grande de las eliminatorias, como ellos calificaron a Brasil, en Sao_Paulo, ante el que _ perdieron 2-3 jugando un buen partido.”

En este caso falla porque en “equipo ecuatoriano” no hay ninguna palabra masculino plural. Cabe destacar que si tuvieramos información más detallada de los sustantivos como ser que un equipo es un conjunto de personas... podríamos considerarlo cuando estamos buscando un plural.

Oración 1035
“ "Es evidente que, aunque Zaplana pone por excusa que el desarrollo urbanístico viene marcado por nuestras leyes, las reformas que ellos han hecho últimamente de las leyes han favorecido el nivel de especulación y la proliferación de masas urbanísticas incontroladas que hacen que el agua sea más escasa y limitada", señaló _.”

Oración 1386
“_ Fue el hombre que rivalizó con Fausto_Coppi y cuyos duelos dividieron a Italia, pero no a ellos, que incluso dieron lecciones al mundo ayudándose, aunque sin dirigirse la palabra.”

En este caso, es un problema de la anotación aparece (39 y luego (39) en “que” (entonces tomo como referente “que”, pero en realidad no habia referente), pero en realidad el pronombre no tiene correferente. El sistema da “duelos” por ser esta palabra masculino plural.

Oración 1407
“Año, siglo y milenio viejos, el 1_de_enero muertos, pero ellos sólo son fechas, números inanes. ”

En este caso el referente es “Año, siglo y milenio viejos”, el algoritmo no lo encuentra por no haber palabra de categoría sustantivo masculino plural entre esas palabras.

Oración 1939
“Llamazares comentó que en su formación política _ sabían que la propuesta de coalición "no era vista de forma favorable" por_parte_del PSOE, al_igual_que ellos no veían "favorablemente su propuesta de retirada de un número de provincias".”

Aca se cometió un error en buscar el correferente que daba la anotación
<palabra id='5' pal='su' p_of_speach='d' morpho='postype=possessive|gen=c|num=s|person=3' head='6' rel_dep='spec' ent='_' corref='(1)|(3'></palabra>
<palabra id='6' pal='formación' p_of_speach='n' morpho='postype=common|gen=f|num=s' head='4' rel_dep='sn' ent='_' corref='_'></palabra>
<palabra id='7' pal='política' p_of_speach='a' morpho='postype=qualificative|gen=f|num=s' head='6' rel_dep='s.a' ent='_' corref='3)'></palabra>
<palabra id='8' pal='_' p_of_speach='p' morpho='_' head='9' rel_dep='suj' ent='_' corref='(3)'></palabra>
<palabra id='9' pal='sabían' p_of_speach='v' morpho='postype=main|num=p|person=3|mood=indicative|tense=imperfect' head='2'

rel_dep='cd' ent='_' corref='_'></palabra>
se dió como referente “su formación politica”, pero el referente debe ser el siguiente que es un “_” elidido. Si se hubiera dado bien la referencia el algoritmo hubiera debuelto true.

Para el pronombre “ellas”

En las siguientes oraciones vemos como la falta de referente hace fallar al algoritmo.

Oración 246
“El dos veces campeón del Abierto_de_Estados_Unidos reiteró que aunque el tenis seguía siendo una de ellas, también existía un elevado grado de confusión en su valoración de lo que puede ser el futuro como profesional. ”

Oración 4126
“Tampoco hay que recordar que en un año tan académico, ellas se han llevado_el_gato_al_agua bailando lo peor que _ sabían, riéndose a la cara de tanta pretendida profesionalidad. ”

Oración 4140
“_ Ha escrito con ellas todas las canciones, excepto Aserejé.”

Oración 4765
“En ellas, _ escribe que Nixon estaba tan "fuera_de_sí" por lo que _ percibía "como otra Cuba" que eso explicaba "su insistencia en hacer cualquier cosa que anulara la negligencia anterior".”

Oración 5562
“Entre ellas, no faltará la canción Forever_Love, de los X-Japan, grupo del que el ministro es fan declarado hasta el punto de que _ sonó con fuerza en todos sus actos de campaña electoral.”

Oración 6155
“"_ Se matriculan en menos materias si _ ven que _ no pueden con ellas y _ van a suspender", comentó el jefe de estudios de la CRUE, Miguel_Ángel_Quintanilla.”

Oración 6186
“En algunas de ellas pueden encontrarse chats (páginas de conversación) donde los adeptos se comunican y un tablón donde dejar anuncios diversos.”

Conclusiones y trabajo futuro

- En lo estudiado se vió que enumeraciones de cosas masculinas singular se corresponden con un pronombre masculino plural, por lo tanto, se podría intentar buscar enumeraciones con o sin conjunciones cuando no se encuentra un sustantivo masculino plural para el pronombre ellos. También hay que tener en cuenta que las enumeraciones o conjunciones mixtas, es decir de femeninos y masculinos también pueden ser nombradas con el pronombre "ellos". Para los femeninos necesitaremos enumeraciones femeninas con o sin conjunciones. Estas distinciones no son aplicables en el inglés.
- Se debería agregar información de cuando una palabra es colectiva ya que representan plurales, los cuales pueden ser referenciados por pronombres plurales (ellos, ellas).
- Agregar preferencias semánticas (en algunos casos fueron agregadas restricciones, o "permisos", como ser puede ser un "sustantivo propio", no puede ser una localidad)
- Agregar restricciones y preferencias según cual es el pronombre estudiado, y su ubicación en la oración, en una frase nominal, en una frase preposicional.
- Distinguir partes del discurso que son aclaraciones, en las cuales no se puede encontrar un referente.
- Hacer otro algoritmo diferente para tratar otras clases de pronombres.
- Comparar los resultados obtenidos con los resultados de alguna herramienta aplicada al mismo corpus.
- Probar con los otros 3 corpus que provee el SemEval (del neerlandés, alemán y el inglés)
- Modificar el algoritmo para que en vez de devolver una palabra saque el subárbol.

Herramientas disponibles

GuiTAR
BART: A Modular Toolkit for Coreference Resolution
Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman,
Alan Jern, Jason Smith, Xiaofeng Yang4, Alessandro Moschitti -2008-
OpenNPL tool
CherryPicker
ANNIE: a Nearly-New Information Extraction System
GATE
JavaRAP -tiene un demo online funciona para inglés
Lymba (no libre)

Bibliografía

Hobbs, Jerry (1978). "Resolving pronoun references." Lingua 44:311-338.
http://www.isi.edu/~hobbs/ResolvingPronounReferences.pdf

An Empirical Approach to Spanish Anaphora Resolution.
Ferrández, A.; Palomar, M.; Moreno, L.
Issn: 0922-6567
Revista: Machine Translation
An Algorithm for Anaphora Resolution in Spanish Text.

Palomar, M.; Ferrández, A.; Moreno, L.; Martínez-Barco, P.; Peral, J.;
Saiz-Noeda, M.; Muñoz, R.
Issn: 0891-2017
Revista: Computational Linguistics
Supervised Models for Coreference Resolution

Altaf Rahman and Vincent Ng
Human Language Technology Research Institute
Proceedings of the 2009 Conference on Empirical Methods in
Natural Language Processing (EMNLP), Suntec, Singapore, August 2009, pp. 968—977.

Joel R. Tetreault: A Corpus-Based Evaluation of Centering and Pronoun Resolution.

M. Strube: Never Look Back: An Alternative to Centering, 1998. (S-list approach)

Marilyn A. Walker: Evaluating Discourse Processing Algorithms.

A. Kehler: Current Theories of Centering for Pronoun Interpretation: A Critical Evaluation.

Documentación del AnCora corpus http://clic.ub.edu/corpus/es/documentacion
Información del catalán http://es.wikipedia.org/wiki/Pronombre_d%C3%A9bil_en_catal
%C3%A1n

GuiTAR http://cswww.essex.ac.uk/Research/nle/GuiTAR/
Annie http://gate.ac.uk/g8/page/print/2/sale/tao/splitch6.html
Gate project http://gate.ac.uk/sale/taln02/taln-ws-coref.pdf
Lymba http://www.lymba.com/technology/coreference.html
OpenNPL tools http://incubator.apache.org/opennlp/
BART http://www.bart-coref.org/
CherryPicker http://www.alterspective.com/cherrypicker/
TextPro http://textpro.fbk.eu/tagset_it.html
MALT Parser http://maltparser.org/
Analizador sintáctico del español y el catalán http://clic.fil.ub.es/demo_sintactico/anali.php