

Elaborazione del linguaggio naturale
a.a. 2006-2007

COREFERENCE RESOLUTION

Alice Tani

Coreference Resolution: che cos'è?

Coreferenza: riferimento di due elementi di un testo o di una frase a uno stesso referente.

es. *“Ho visto l'uva e la ho mangiata.”*

Coreference Resolution: processo di determinare se due espressioni in linguaggio naturale fanno riferimento alla stessa entità nel mondo.

Coreference Resolution: che cos'è?

(2)

- Task fondamentale per l'analisi e la comprensione di un testo nei sistemi di elaborazione del linguaggio naturale, in particolare in sistemi di *Information Extraction* e di *Question Answering*.
- Se ne occupava già la *Message Understanding Conference (MUC)* alla fine degli anni '90.
- Attualmente è anche un task interno al *Programma ACE (Automatic Content Extraction)*.

Coreference relations

- MUC definì uno schema di annotazione che consentisse di evidenziare le *relazioni di coreferenza* all'interno di un testo, ossia che permettesse di “marcare” e collegare insieme tutte le espressioni che si riferiscono alla stessa entità.
- Una relazione di coreferenza è definita da MUC come una *relazione di identità*, IDENTITY o IDENT, tra due elementi testuali denominati *markables*.

es. *Ho visto* <COREF ID = “1”>l'uva</COREF> e <COREF ID = “2” TYPE = “IDENT” REF = “1”>la</COREF> *ho mangiata.*

“l'uva” e “la” sono i markables coinvolti nella relazione di coreferenza.

Coreference relations (2)

- La relazione di coreferenza è *simmetrica*: se A è coreferenziale con B, allora B è coreferenziale con A.
- La relazione di coreferenza è *transitiva*: se A è coreferenziale con B e B è coreferenziale con C, allora A è coreferenziale con C.
- Queste proprietà producono un insieme di *classi di equivalenza* tra gli elementi marcati, tali che:
 - ogni elemento è contenuto in una sola classe;
 - tutti gli elementi di una stessa classe sono coreferenziali.

Coreference relations (3)

- Pertanto, gli elementi di una stessa classe creano all'interno del testo annotato una *catena di coreferenza*.
- Se tre markables, A, B e C, sono coreferenziali, la relazione può essere annotata nel testo in vari modi. Per esempio:
 - con un puntatore REF sia in B sia in C che punta ad A, oppure
 - con un puntatore REF in B che punta ad A e un puntatore REF in C che punta a B.

Coreference relations: problemi

La natura della relazione di coreferenza come insiemi di classi di equivalenza produce un problema nel caso di espressioni che indicano un **cambiamento nel tempo**.

es. *“La temperatura passò da 20 gradi a 25 gradi”*

“**La temperatura**” è inizialmente coreferenziale con “**20 gradi**” e in seguito con “**25 gradi**”.

“20 gradi” e “25 gradi” non possono appartenere alla stessa classe di equivalenza!

Soluzione proposta: annotare la relazione di coreferenza solamente tra “La temperatura” e il suo valore più recente, “25 gradi”.

Markables

- Gli elementi annotati coinvolti in una relazione di coreferenza possono essere:
 - *sintagmi nominali* (NOUN PHRASES);
 - *pronomi* (dimostrativi, personali, possessivi);
 - *nomi propri*.
- Un NP è un markable quando è oggetto di:
 - un'affermazione (es. *I have a dog*);
 - una negazione (es. *I don't have a dog*);
 - una domanda (es. *Do you have a dog?*).

Markables (2)

- Sono annotati come markables anche:
 - gli NP che fanno parte di *sintagmi preposizionali* (es. *la riduzione del salario*);
 - le *coniunzioni* (es. *The sleepy boys and girls enjoy their breakfast*);
 - i *membri individuali delle coniunzioni*, se sono coreferenziali separatamente con altri markables (es. *Edna Fribble and Sam Morton addressed the meeting yesterday. Ms. Fribble discussed coreference, Mr. Morton discussed named entities.*);
 - le *apposizioni* (es. *Julius Caesar, the well-known emperor*).

Markables (3)

- Non sono considerati markables:
 - i *pronomi interrogativi* (es. *Who is your boss?* oppure *Which car do you have?*)
 - le *proposizioni*.
- La relazione di coreferenza è annotata *solamente* tra coppie di elementi entrambi i quali sono markables.

es. “*Non è una brava cuoca, questo è certo!*”

“*questo*” non è considerato un markable perché il suo antecedente “*Non è una brava cuoca*” è una *proposizione*.

Coreference Resolution: tecniche

Metodi utilizzati:

- *Corpus-based Machine Learning*
- *Knowledge-based*
- *Approcci ibridi*

Un approccio Machine Learning

- Richiede un *corpus di documenti di “training”* annotato con le catene di coreferenza.
- In un documento di training, per ogni coppia di markables viene generata un'*istanza di training* nella forma di *vettore di features*.
- Un'istanza di training può essere *positiva* (se entrambi i markables appartengono alla stessa catena di coreferenza) o *negativa*.
- Le istanze di training sono date a un algoritmo di apprendimento per costruire un *classificatore*.

Un approccio Machine Learning (2)

Per determinare le catene di coreferenza in un nuovo testo, un approccio Machine Learning standard decompone il task in due passi:

- **Classificazione:** ogni potenziale coppia di markables (*candidato antecedente, anafora*) è presentata al classificatore, il quale decide se i due markables sono coreferenziali oppure no.
- **Clustering:** le coppie identificate come coreferenziali al primo passo sono “fuse” tra loro per formare le catene di coreferenza.

Determinare i markables

- *Prerequisito* per la Coreference Resolution: determinare la maggior parte (se non tutti!) dei markables in un nuovo testo.
- Si utilizza una *pipeline* di moduli di elaborazione del linguaggio naturale (NLP).
- *Scopo* dei moduli NLP:
 - definire l'*estensione* dei markables;
 - raccogliere le informazioni necessarie per la successiva generazione delle *features* nelle istanze di training.

Pipeline di moduli NLP

1. *Tokenization* e divisione in *frasi*;
2. Elaborazione *morfologica*;
3. *Tagging*;
4. Identificazione *NP*;
5. *Named Entity Recognition*;
6. Fusione tra le NP identificate al passo 4 e le NE riconosciute al passo 5: se un NP si sovrappone con una NE, l'estensione del NP è “aggiustata” per includere la NE.

Pipeline di moduli NLP (2)

7. Estrazione *NP annidati*; ad esempio:

- NP annidati in NP possessivi:
 - “*his long-range strategy*”
 - “*Eastern's father*”
- modificatori:
 - “*wage reductions*”
 - “*Union representatives*”

8. Determinazione della *classe semantica* per i markables che non hanno un tipo entità.

- I markables per la Coreference Resolution sono l'*unione* delle entità, degli NP e degli NP annidati trovati.

Determinare i vettori di features

- Per costruire un motore di Coreference Resolution learning-based è utile definire un insieme di features che aiuti a determinare se due markables sono coreferenziali oppure no.
- Tipi diversi di NP possono avere un comportamento molto diverso rispetto al modo in cui coreferiscono e pertanto richiedere strategie di Coreference Resolution diverse.
- Vedremo in seguito che, in generale, un approccio Machine Learning non è mai puramente corpus-based: ci sono molte fonti di conoscenza assai utili che possono essere sfruttate per identificare una relazione di coreferenza.

Esempi di features

- Sia (i, j) una coppia di markables, dove j è l'anafora e i il potenziale antecedente. Esempi di possibili features sono:
 - **Distance (DIST)**: distanza tra i e j in termini di frasi (0 se i e j sono nella stessa frase);
 - **i-Pronoun (I_PRONOUN) / j-Pronoun (J_PRONOUN)**: indica se i (j) è un pronome;
 - **String Match (STR_MATCH)**: indica se la stringa di i fa match con la stringa di j . Eventuali articoli e pronomi dimostrativi sono rimossi prima di eseguire il paragone. Es. *the license* fa match con *this license*; *that computer* fa match con *computer*;

Esempi di features (2)

- **Definite Noun Phrase (DEF_NP)**: indica se j è un NP definito (inizia con *the*). Es. *the car*;
- **Demonstrative Noun Phrase (DEM_NP)**: indica se j è un NP dimostrativo (inizia con *this*, *that*, *these* o *those*);
- **Number Agreement (NUMBER)**: indica se i e j hanno lo stesso numero (entrambi singolari o entrambi plurali);
- **Semantic Class Agreement (SEMCLASS)**: indica se i e j hanno la stessa classe semantica. Es. “male”, “female”, “person”, “organization”, “location”, “date”, “time”, “money”, “percent”, “object”;

Esempi di features (3)

- **Gender Agreement (GENDER)**: indica se i e j hanno lo stesso genere;
- **Both-Propor-Names (PROPER_NAME)**: indica se i e j sono entrambi nomi propri;
- **Alias (ALIAS)**: indica se i e j sono alias, ossia si riferiscono alla stessa entità. Es. *01-08* e *January 8*; *Mr. Simpson* e *Homer Simpson*; *IBM* e *International Business Machine Corp*;
- **Appositive (APPOSITIVE)**: indica se j è in apposizione a i . Es. *Bill Gates, the chairman of Microsoft*.

Generare le istanze di training

- Sia $A1-A2-A3-A4$ una catena di coreferenza annotata in un documento di training.
- Metodo possibile per generare le istanze di training *positive*: usare le coppie di markables immediatamente *adiacenti* nella catena ($A1-A2$, $A2-A3$, $A3-A4$).
- Tra i due membri di ciascuna coppia (*antecedente, anafora*) possono essere presenti altri markables che non appartengono a nessuna catena di coreferenza oppure appartengono ad un'altra catena. Ciascuno di essi è accoppiato con l'anafora per generare un'istanza di training *negativa*.

Esempio: se i markables a , b e $B1$ appaiono tra $A1$ e $A2$, le istanze negative saranno $a-A2$, $b-A2$ e $B1-A2$.

Generare le istanze di training (2)

- *Prerequisito*: un elemento annotato in un documento di training come appartenente a una catena di coreferenza *deve* essere identificato come markable dalla pipeline di moduli NLP *prima* di poter essere usato per la generazione di un istanza di training.
- *Motivo*: le informazioni necessarie per la creazione del *vettore di features* che caratterizza ciascuna istanza sono raccolte dalla pipeline di moduli NLP.
- Se un NP annotato non viene identificato come markable, non potrà contribuire alla formazione di nessuna istanza di training.

Generare le istanze di training: esempio

Esempio di testo annotato con catene di coreferenza:

(Eastern Air)_{a1} Proposes (Date For Talks on ((Pay)_{c1}-Cut)_{d1} Plan)_{b1}.
(Eastern Airlines)_{a2} executives noticed (union)_{e1} leaders that the carrier wishes to discuss selective ((wage)_{c2} reductions)_{d2} on (Feb. 3)_{b2}.
((Union)_{e2} representatives who could be reached)_{f1} said (they)_{f2} hadn't decided whether (they)_{f3} would respond.
By proposing (a meeting date)_{b3}, (Eastern)_{a3} moved one step closer toward reopening current high-cost contract agreements with ((its)_{a4} unions)_{e3}.

Generare le istanze di training: esempio (2)

Esempio di markables identificati dalla pipeline di moduli NLP:

(Eastern Air)₁ Proposes (Date)₂ For (Talks)₃ on (Pay-Cut Plan)₄.
((Eastern Airlines)₅ executives)₆ noticed ((union)₇ leaders)₈ that (the carrier)₉ wishes to discuss (selective (wage)₁₀ reductions)₁₁ on (Feb. 3)₁₂.
((Union)₁₃ representatives)₁₄ who could be reached said (they)₁₅ hadn't decided whether (they)₁₆ would respond.
By proposing (a meeting date)₁₇, (Eastern)₁₈ moved (one step)₁₉ closer toward reopening (current high-cost contract agreements)₂₀ with ((its)₂₁ unions)₂₂.

Generare le istanze di training: esempio (3)

- Consideriamo la catena *e* nel testo annotato:

*(union)*_{e1} – *(Union)*_{e2} - *(its unions)*_{e3}

La pipeline ha identificato tutti i markables corrispondenti:

*(union)*₇ – *(Union)*₁₃ - *(its unions)*₂₂

Le istanze *positive* risultanti sono: *((Union)*₁₃, *(its unions)*₂₂) e *((union)*₇, *(Union)*₁₃).

Le istanze *negative* tra *(union)*₇ e *(Union)*₁₃ sono: *((the carrier)*₉, *(Union)*₁₃), *((wage)*₁₀, *(Union)*₁₃), *((selective wage reductions)*₁₁, *(Union)*₁₃) e *((Feb. 3)*₁₂, *(Union)*₁₃).

In modo analogo sono generate le istanze negative tra *(Union)*₁₃ e *(its unions)*₂₂.

Generare le istanze di training: esempio (4)

- Consideriamo invece la catena *d* nel testo annotato:

*(Pay-Cut)*_{d1} – *(wage reductions)*_{d2}

La pipeline ha identificato i markables:

(Pay-Cut Plan)₄ e (selective (wage)₁₀ reductions)₁₁

Nessuno dei markables estratti dalla pipeline corrisponde a uno degli elementi annotati nel documento di training.

In questo caso, nessuna istanza di training positiva o negativa viene generata per gli elementi annotati come appartenenti alla catena *d*.

Costruire un classificatore

- I vettori di features che rappresentano le istanze, positive e negative, generate dai documenti di training sono sottomesse a un algoritmo di apprendimento per costruire un *classificatore*.
- Vari tipi di classificatori possibili.
- *Un esempio*: classificatore strutturato ad *albero di decisione*.
 - Ogni nodo dell'albero è una domanda relativa a una delle features.
 - In base alla risposta, viene determinato il nuovo cammino da seguire.
 - Quando viene raggiunta una foglia, viene restituita l'etichetta corrispondente (+ se i due markables della coppia coreferiscono, - altrimenti).

Costruire un classificatore: esempio di struttura ad albero

```
if (STR_MATCH = +) return +;
if (STR_MATCH = -)
    if (J_PRONOUN = -)
        if (APPOSITIVE = +) return +;
        if (APPOSITIVE = -)
            if (ALIAS = +) return +;
            if (ALIAS = -) return -;
    if (J_PRONOUN = +)
        if (GENDER = 0) return -; // false
        if (GENDER = 2) return -; // unknown
        if (GENDER = 1) // true
            if (I_PRONOUN = +) return +;
            if (I_PRONOUN = -)
                if (DIST > 0) return -;
                if (DIST <= 0)
                    if (NUMBER = +) return +;
                    if (NUMBER = -) return -;
```

Generare catene di coreferenza in un nuovo testo

- *Prima* di determinare le catene di coreferenza occorre estrarre dal documento tutti i possibili markables.
- Ogni markable è una possibile *anafora* e ogni markable che precede l'anafora è un possibile *antecedente* dell'anafora.
- Se l'anafora è un markable *annidato*, i suoi possibili antecedenti *non* possono essere markables che hanno la stessa radice dell'anafora. Però sono antecedenti possibili altri markables radice e i loro figli che precedono l'anafora nel documento.

Es. *Mr. Tom's daughter* e *His daughter's eyes* sono due markables radice che appaiono in quest'ordine nel documento.

Possibili antecedenti di *His non* possono essere *His daughter* o *His daughter's eyes*, **ma** possono esserlo *Mr. Tom* o *Mr. Tom's daughter*.

Generare catene di coreferenza in un nuovo testo (2)

L'algoritmo di *Coreference Resolution*:

- considera ogni markable j a partire dal secondo markable del documento come una possibile *anafora*;
- per ogni j , considera ogni markable i che precede j come un potenziale *antecedente*;
- per ogni coppia (i,j) genera un *vettore di features* e lo sottomette al *classificatore*;
- se il classificatore restituisce true, aggiunge l'anafora j alla catena dell'antecedente i e passa a considerare la prossima anafora.

I potenziali antecedenti sono considerati a partire da quello immediatamente precedente all'anafora e procedendo all'indietro nell'ordine inverso del documento, fino a che viene trovato un antecedente o non ci sono più markables da testare.

Un esempio

(**Ms. Washington**)₇₃'s candidacy is being championed by (several powerful lawmakers)₇₄ including ((**her**)₇₆ boss)₇₅, (Chairman John Dingell)₇₇ (D., (Mich.)₇₈) of (the House Energy and Commerce Committee)₇₉. (**She**)₈₀ currently is (a counsel)₈₁ to (the committee)₈₂. (**Ms. Washington**)₈₃ and (Mr. Dingell)₈₄ have been considered (allies)₈₅ of (the (securities)₈₇ exchanges)₈₆, while (banks)₈₈ and ((futures)₉₀ exchanges)₈₉ have often fought with (them)₉₁.

La tabella mostra alcune coppie di markables analizzate dal classificatore per generare la catena:

Ms. Washington – her – She - Ms. Washington

Un esempio (2)

Antecedent	Anaphor	Feature Vector	Corefers?
(several powerful lawmakers) ₇₄	(her) ₇₆	0,1,-2,-,-,+,-,-,-,-,-	No
(Ms. Washington) ₇₃	(her) ₇₆	0,1,+1,-,-,+,-,-,-,-,-	Yes
(the House Energy and Commerce Committee) ₇₉	(She) ₈₀	1,0,+0,-,-,+,-,-,-,-,-	No
(Mich.) ₇₈	(She) ₈₀	2,0,+0,-,-,+,-,-,-,-,-	No
(Chairman John Dingell) ₇₇	(She) ₈₀	3,1,+0,-,-,+,-,-,-,-,-	No
(her) ₇₆	(She) ₈₀	3,1,+1,-,-,+,-,-,-,-,+	Yes
(the committee) ₈₂	(Ms. Washington) ₈₃	1,0,+0,-,-,-,-,-,-,-,-	No
(a counsel) ₈₁	(Ms. Washington) ₈₃	1,1,+2,-,-,-,-,-,-,-,-	No
(She) ₈₀	(Ms. Washington) ₈₃	1,1,+1,-,-,-,-,-,-,-,+	No
(the House Energy and Commerce Committee) ₇₉	(Ms. Washington) ₈₃	2,0,+0,+,-,-,-,-,-,-,-	No
(Mich.) ₇₈	(Ms. Washington) ₈₃	3,0,+0,+,-,-,-,-,-,-,-	No
(Chairman John Dingell) ₇₇	(Ms. Washington) ₈₃	4,1,+0,+,-,-,-,-,-,-,-	No
(her) ₇₆	(Ms. Washington) ₈₃	4,1,+1,-,-,-,-,-,-,-,+	No
(her boss) ₇₅	(Ms. Washington) ₈₃	4,1,-0,-,-,-,-,-,-,-,-	No
(several powerful lawmakers) ₇₄	(Ms. Washington) ₈₃	4,1,-2,-,-,-,-,-,-,-,-	No
(Ms. Washington) ₇₃	(Ms. Washington) ₈₃	4,1,+1,+,-,-,-,-,+,-,-	Yes

Un approccio Knowledge-based

- Un esempio di approccio Knowledge-based per la *lingua giapponese*.
- In Giapponese, i pronomi *non* sono molto usati: di solito un'anafora è rappresentata da un *nome comune* o un *nome proprio*.
- Utile la tecnica della comparazione di stringhe (*string matching*): spesso un'anafora e il suo antecedente condividono stringhe.
- Esistono casi, però, in cui espressioni coreferenziali *non* hanno nessuna stringa in comune: in questi casi, la tecnica *string matching* non può essere applicata.

Un approccio Knowledge-based (2)

- Per riconoscere una relazione di coreferenza tra espressioni assolutamente diverse è assai utile una fonte di conoscenza specifica: la *conoscenza dei sinonimi*.
- Necessario acquisire conoscenza dei sinonimi in anticipo.
- *Risorse* utilizzate per l'estrazione di sinonimi:
 - *corpus liberi* (non annotati): utili per l'estrazione di sinonimi *non comuni*;
 - *definizioni del dizionario*: per sinonimi *comuni*.

Estrazione di sinonimi da corpus liberi

- Quando un sinonimo *non comune* è usato per la prima volta in un testo, spesso l'informazione è scritta utilizzando una *coppia di parentesi*.
 - **Es.** *Suzuki Chosen Hantoo Enerugii Kaihatu Kiko (KEDO) taishi-ga yutai-shita.*
 - **Traduzione:** L'ambasciatore (taishi-ga) Suzuki dell'Organizzazione (Kiko) per lo Sviluppo (Kaihatu) dell'Energia (Enerugii) nella Penisola (Hantoo) Coreana (Chosen) (KEDO) si è ritirato (yutai-shita).
 - La parola tra parentesi, “*KEDO*”, è un sinonimo per “*Chosen Hantoo Enerugii Kaihatu Kiko*”.
- Le parentesi non sono usate sempre per indicare un sinonimo. Per esempio, possono indicare un attributo di un nome, come l'età o l'affiliazione.

Estrazione di sinonimi da corpus liberi (2)

- *Problema*: estrarre solo le coppie di parentesi che indicano sinonimi.
- *Assunzione*: se una coppia A, B che appare in un'espressione del tipo “ $A (B)$ ” è una coppia di sinonimi, allora la *frequenza* delle espressioni tra parentesi è alta e la coppia inversa “ $B (A)$ ” può a sua volta comparire nel corpus.
- *Metodo* per estrarre coppie di sinonimi da espressioni tra parentesi:
 - si conta la *frequenza* di coppie A, B tali che B è un'espressione tra parentesi e A è il nome che la precede, cioè “ $A (B)$ ”;
 - se la frequenza supera una certa *soglia* prestabilita, la coppia A, B è giudicata come una coppia di sinonimi.

Estrazione di sinonimi da corpus liberi (3)

- Soglie di frequenza diverse sono impostate per tipi diversi di coppie di sinonimi.
- *Esempi* di coppie di sinonimi:
 - un'espressione è composta da lettere dell'alfabeto inglese e l'altra no;
 - un'espressione è composta da lettere dell'alfabeto giapponese KATAKANA e l'altra no;
 - un'espressione è composta da caratteri cinesi e l'altra è la sua abbreviazione.

Estrazione di sinonimi da dizionario

- Si utilizzano le definizioni di un dizionario per estrarre sinonimi *comuni*. **Es.** “*USA*” e “*America*”.
- *Metodo* di estrazione: per ogni lemma *A* del dizionario:
 - se la definizione termina con “*no ryaku*” (*abbreviazione di*) o “*no koto*” (*sinonimo di*), si estrae il resto della definizione come un potenziale sinonimo *B*; altrimenti, si estrae l'intera definizione da considerare come *B*;
 - se anche *B* è un lemma di dizionario, la coppia *A,B* è estratta come coppia di sinonimi.

Strategia per la Coreference Resolution

1. Si identificano tutti i markables presenti nel testo.
2. Per ogni anafora:
 - dalla posizione dell'anafora fino all'inizio del testo, si considera ogni markable come un possibile antecedente;
 - se l'anafora e il potenziale antecedente hanno un completo *string matching* oppure l'anafora è un *sinonimo* dell'antecedente, si classifica la coppia (antecedente, anafora) come coreferenziale e si passa a considerare la prossima anafora.

Verso un approccio ibrido

- In un approccio Machine Learning, esistono varie possibilità per generare le istanze di training da un corpus annotato.
- *Metodo più semplice*: ogni markable è collegato con *tutti* i markables che lo precedono per formare un insieme di coppie; una coppia è classificata come positiva se entrambi i markables appartengono alla stessa catena di coreferenza secondo l'annotazione del corpus, negativa altrimenti.
- Problemi con l'approccio “*generare tutte le coppie possibili*”...

Verso un approccio ibrido (2)

- *Distribuzione non equilibrata*: il numero di istanze positive generate è troppo piccolo rispetto al numero di istanze negative.
- Alcune istanze positive potrebbero essere *irrilevanti* o troppo *difficili*.
- Un'istanza è *irrilevante* se la sua eliminazione dall'insieme di istanze di training non comporta una perdita di informazione.

Es. “*This deal means that Bernard Schwartz can focus most of his time on Globalstar*” said Robert Kaimovitz. *Schwartz said Monday that... [...]*”

Se abbiamo le istanze (*Bernard Schwartz-his*) e (*Bernard Schwartz-Schwartz*), l'istanza (*his-Schwartz*) è irrilevante.

Un solo antecedente è sufficiente per risolvere un'anafora.

Verso un approccio ibrido (3)

- Un'istanza è troppo *difficile* se non esiste uno schema definito per identificare la relazione di coreferenza corrispondente, oppure esiste ma è troppo sofisticato.

Es. *Metonimia*: nomi di nazioni o di loro capitali utilizzati per indicare i rispettivi governi.

“Beijing remained committed in talks with ... [...] China would not impose sanctions on Taiwanese goods.”

- Istanze troppo difficili possono introdurre confusione nel sistema.

Es. *“This deal means that Bernard Schwartz can focus most of his time on Globalstar” said Robert Kaimovitz. Schwartz said Monday that... [...]”*

Mantenere l'istanza (*his-Schwartz*) può far apprendere al classificatore regole sbagliate che lo portano a non saper distinguere tra istanze positive come (*his-Schwartz*) da quelle negative come (*his-Robert Kaimovitz*).

Verso un approccio ibrido (4)

- *Obiettivo*: avere un componente extra che permetta di:
 - ridurre le istanze negative non necessarie;
 - individuare ed eliminare le istanze positive irrilevanti o troppo difficili.
- Metodi *corpus-based*: non utilizzano nessuna fonte di conoscenza linguistica e dipendono dalle caratteristiche dello specifico corpus utilizzato.
- Metodi *knowledge-based*: si basano su proprietà linguistiche dei markables per identificare le istanze più rilevanti, indipendentemente dalle caratteristiche del corpus.

Selezione di un campione

- Metodo *corpus-based* per ridurre le istanze negative non necessarie: scartare ogni coppia (i,j) in cui i compare nel testo *prima* del *più lontano* antecedente di j che è stato identificato.

Es. “*This deal means that Bernard Schwartz can focus most of his time on Globalstar*” said Robert Kaimovitz. *Schwartz said Monday that... [...]*”

L'istanza $(This\ deal, his)$ può essere scartata, così come $(This\ deal, Schwartz)$.
- *Intuizione*: un solo antecedente è sufficiente per collegare un'anafora alla sua catena di coreferenza. Una volta identificato tale antecedente, i markables precedenti non vengono elaborati. Pertanto, le coppie corrispondenti non portano alcuna nuova informazione e possono essere scartate.

Selezione di un campione (2)

- Metodo *knowledge-based* per selezionare un campione di istanze rilevanti.
- *Intuizione*: tipi diversi di anafore richiedono strategie di risoluzione diverse; un antecedente può essere considerato rilevante per un tipo di anafora, ma irrilevante per un altro tipo.
- *Sistema ibrido*: combina l'approccio Machine Learning con l'approccio Knowledge-based.
- *Approccio utilizzato*: applicazione di *filtri linguisticamente motivati* alle istanze di training *prima* di sottometerle al classificatore.

Tipi di anafore

- Facciamo distinzione tra vari tipi di anafore:
 - *pronomi*;
 - *nomi propri*;
 - *NP definiti*;
 - *NP indefiniti*.
- Ogni tipo ha un comportamento linguistico diverso rispetto alla relazione di coreferenza.

Pronomi

- Fattore importante: *distanza* tra un pronome e un suo potenziale antecedente; di solito i pronomi hanno i loro antecedenti nella stessa frase o nelle frasi immediatamente precedenti.
- *Strategia*: escludere tutte le coppie in cui l'antecedente non è sufficientemente vicino all'anafora pronominale.
- *Criteri possibili*:
 - la distanza massima tra antecedente e anafora è di 2 frasi;
 - la distanza massima tra antecedente e anafora è di 5 frasi;
 - antecedente e anafora sono nello stesso paragrafo;
 - l'antecedente è il più vicino possibile all'anafora.

Nomi propri

- Di solito, un'entità (NE) è introdotta nel discorso con un nome proprio e, solo in seguito, è indicata con varie descrizioni (NP definiti, pronomi...).
- Dal punto di vista della Coreference Resolution, questo vuol dire che di solito una NE è una prima menzione oppure ha almeno un antecedente che è anch'esso una NE.
- *Strategia*: considerare solo le coppie in cui anche l'antecedente è un nome proprio e insegnare al classificatore a identificare varianti dello stesso nome.
 - *Es. Homer Simpson e Mr. Simpson*

NP definiti

- Se un NP definito ha un potenziale antecedente con la stessa *testa sintattica*, è sufficiente confrontare le strutture dell'antecedente e dell'anafora (in particolare i *modificatori*, i *sintagmi preposizionali* e le *proposizioni relative*).

Es. (the lazy boy, the boy)

- Se un NP è una nuova descrizione di un oggetto già menzionato, occorre un'informazione semantica sulle teste nominali dell'antecedente e dell'anafora.

Es. (SGI's technology, the software) oppure (the country, the nation)

- *Strategia*: includere solo le coppie in cui l'antecedente ha la stessa testa dell'anafora; se non esistono coppie siffatte, considerare tutte le coppie in cui la testa sintattica dell'antecedente non è un pronome.

NP indefiniti

- Di solito gli NP indefiniti *non* svolgono il ruolo di anafora.
- Tuttavia un NP indefinito può appartenere a una catena di coreferenza quando è il predicato in una *costruzione appositiva*.
 - **Es.** “*Peter VanHausen, a Manchester banker*”
- Un NP può essere anaforico anche per ragioni “oscure”; per esempio, è un NP indefinito usato come prima menzione in un discorso indiretto, ma che fa riferimento a un oggetto già introdotto nel testo principale:
 - “*The **craft**'s distress had lasted long enough for one driver to call a radio station to report that a **helicopter** was in trouble.*”
 - Meglio non includere queste istanze “oscure”, poiché portano confusione nel sistema.

Filtri positivi e negativi

- In un *approccio ibrido*:
 - dopo la generazione delle istanze di training, ogni istanza viene “filtrata” da uno o più filtri di regole;
 - ogni istanza che rispetta la regola del filtro è classificata direttamente dal filtro stesso;
 - le istanze rimanenti sono sottoposte al classificatore.
- Filtri *negativi* riducono il numero di istanze negative.
- Filtri *positivi* selezionano un campione di istanze positive.
- Un filtro può essere applicato da solo oppure in combinazione con altri filtri.

Filtri positivi e negativi: un esempio

- **fdef**: filtra tutte le istanze in cui l'anafora è un NP indefinito e assegna un'etichetta negativa a queste istanze;
- **fhead**: filtra tutte le istanze in cui anafora e antecedente sono più distanti di tre frasi l'uno dall'altra se non condividono la stessa testa sintattica;
- **fagree**: si applica solo alle coppie in cui l'anafora è un pronome; filtra tutte le istanze in cui l'antecedente non ha lo stesso numero o lo stesso genere dell'anafora;
- **fmatch**: assegna un'etichetta positiva a tutte le istanze in cui anafora e antecedente hanno un completo “string match”; tutte le altre istanze che contengono la stessa anafora con un altro antecedente ottengono un'etichetta negativa.
- **f3s**: filtra tutte le coppie (antecedente, pronome) in cui l'antecedente si trova a una distanza maggiore di tre frasi dal pronome.

Migliorare il processo di risoluzione

- I filtri linguistici possono essere applicati anche prima di eseguire l'algoritmo di Coreference Resolution per determinare le catene di coreferenza in un nuovo testo.
- *Obiettivo*: migliorare il processo di risoluzione rimuovendo le coppie (i, j) in cui i non è ritenuto un antecedente appropriato per l'anafora j .
- Le coppie rimaste dopo l'applicazione dei filtri sono sottomesse al “giudizio” del classificatore.
- *Esempio*: usare conoscenze sintattiche per identificare gli NP indefiniti che non sono referenziali e rimuoverli affinché non vengano considerati durante il procedimento di ricerca di un antecedente.

Eliminare NP non referenziali

- *Intuizione*: in alcuni contesti, un NP indefinito non presuppone l'esistenza di un'entità individuale, pertanto l'entità a cui l'NP fa riferimento non sarà *mai* menzionata fuori dal contesto locale.

Es. “*Pat is a great teacher*”

“*Pat*” fa riferimento a una specifica entità, ma “*a great teacher*” no. Pertanto “*a great teacher*” *non* è referenziale.

- E' possibile applicare un filtro che etichetti come **NON-REF** gli NP non referenziali *prima* di eseguire l'algoritmo di Coreference Resolution: gli NP etichettati come NON-REF non saranno considerati come potenziali antecedenti.

Contesti in cui un NP indefinito è non referenziale

- *Predicati nominali*: “*Bill is a dog*”;
- *Negazioni*: “*Bill does not have a car*”;
- *Verbi modali*: “*Bill can make a kite*”, “*Bill should make a kite*”;
- *Verbi che implicano una negazione* (presuppongono la non esistenza dell'entità oggetto): “*Bill failed to write a letter*”, “*Bill forgot to write a letter*”;
- *NP negati*: “*no man*”, “*no money*”;
- *Predicati di una costruzione appositiva*: “*Peter VanHausen, a Manchester banker*”;
- *Altri contesti*:
 - “*Pierre will join the board as a nonexecutive director*”;
 - “*Chris failed to find an apartment but I found one*” (“*an apartment*” è non referenziale perché l'anafora “*one*” fa riferimento ad una categoria e non ad una specifica entità).

Un filtro per NP non referenziali

NON-REF if (appositive is true
or determiner is “no”
or numeric-value is true
or (the NP is indefinite
and (negated is true
or predicate is true
or modal is true
or modifier is true)
)
)

Conclusioni

- La *Coreference Resolution* si occupa di riconoscere e raggruppare tutte le espressioni che fanno riferimento a una stessa entità.
- Gli approcci *Learning-based* si basano su un corpus annotato per generare esempi positivi e negativi da “insegnare” a un *classificatore*.
- Gli approcci *Knowledge-based* risolvono le relazioni di coreferenza utilizzando un insieme di fonti di *conoscenza linguistica*.
- Di solito, è utilizzato un *approccio ibrido*: l'approccio Knowledge-based è combinato con l'approccio Machine Learning mediante l'applicazione di *filtri linguistici* ai documenti prima dell'utilizzo del classificatore.
- I filtri possono essere applicati sia ai documenti annotati per l'“allenamento” del classificatore, sia ai documenti in cui devono essere identificate le nuove relazioni di coreferenza.

Bibliografia

- MUC - 7 Coreference Task Definition, 1997.

http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

- ACE – Automatic Content Extraction

<http://www.nist.gov/speech/tests/ace/>

- Wee Meng Soon, Hwee Tou Ng, Daniel Chung Yong Lim, 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases.

<http://acl.ldc.upenn.edu/J/J01-4004.pdf>

- Ryohei Sasano, Daisuke Kawahara, Sadao Kurohashi, 2007. Improving Coreference Resolution Using Bridging Reference Resolution and Automatically Acquired Synonyms.

<http://www-nagao.kuee.kyoto-u.ac.jp/~ryohei/paper/DAARC07.pdf>

Bibliografia (2)

- Olga Uryupina, 2004. Linguistically Motivated Sample Selection for Coreference Resolution.

http://www.coli.uni-saarland.de/~ourioupi/ury_daarc_fin.pdf

- Iris Hendrickx, Veronique Hoste, Walter Daelemans, 2007. Evaluating hybrid versus data-driven coreference resolution.

http://www.cnts.ua.ac.be/Publications/2007/HHD07/daarc_full.pdf

- Donna K. Byron, Whitney Gegg-Harrison, 2004. Eliminating Non-Referring Noun Phrases from Coreference Resolution.

<http://www.cse.ohio-state.edu/~dbyron/DAARC2004-Byron-Geggharr-app.pdf>