



**EVALITA 2011**

*Evaluation of NLP and Speech Tools for Italian*

# Dynamic Threshold for Clustering Person Names

R. Zanoli, F. Corcoglioniti, C. Girardi






# CDC

Home > Racing > MotoGP > News

## MotoGP

15 Aug 2010

### VALENTINO ROSSI AND DUCATI TOGETHER FROM 2011



Ducati and Valentino Rossi have signed a two year agreement for the nine-time World Champion to race with the "Rossa" of Borgo Panigale in the Ducati Team from 2011.

- Two names ***Valentino Rossi***
- One Person

Valentino Rossi faces six months out after breaking leg

Italy MotoGP, Mugello

Date: Saturday 5 June/Sunday 6 June

Saturday BBC coverage: Qualifying - 1150-1500, BBC Red Button/online

Sunday BBC coverage: 125cc and Moto2 races - 0950-1205, BBC Red Button/online; Race live - 1230-1400, BBC Two/online; MotoGP Extra - 1400-1430, BBC Red Button/online



SEE ALSO

- ▶ Matt Roberts' Iv 03 Jun 10 | Mo
- ▶ Steve Parrish's 03 Jun 10 | Mo
- ▶ Lorenzo triumph 23 May 10 | Mc
- ▶ Rossi injured in 16 Apr 10 | Mo
- ▶ MotoGP calend 20 Apr 10 | Mo
- ▶ MotoGP on the 11 Aug 10 | Mc
- ▶ MotoGP standir 16 Aug 09 | Mc
- ▶ Contact the mo 06 Mar 07 | Mc

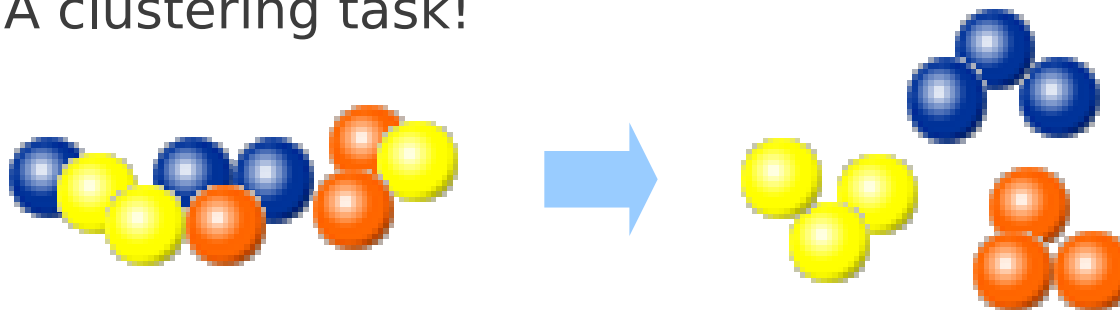
Cross-document coreference (CDC) occurs when the same person, place, or concept is discussed in more than one text source.



# What CDC is not

Different from word sense disambiguation

- The set of persons for a given name (i.e. the senses) is not known a priori.
- A clustering task!





# A general Approach

- Names to be clustered are represented with their context
  - context can be of different sizes: window of words centered around a name, sentence containing name, group of sentences, or even whole document.
  - modeling context can be done in many different ways: bag of words, set of phrases, set of entities, set of relations, etc.



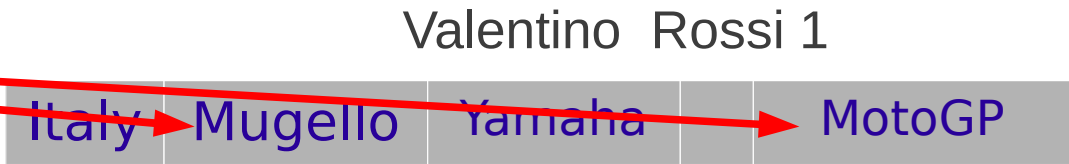
# A general Approach

e.g.

Valentino Rossi faces six months out after breaking leg

MotoGP OFF ALSO  
 Date: 15 June/Sunday 09:00  
 Saturday BBC coverage: Qualifying - 1150-1500, BBC Red Button/online  
 Sunday BBC coverage: 125cc and Moto2 races - 0950-1205, BBC Red Button/online; Race live  
 - 1230-1400, BBC Two/online; MotoGP Extra - 1400-1430, BBC Red Button/online

- Matt Baker's TV 03 Jun 10 | Mo
- Steve Parrish's 03 Jun 10 | Mo
- Lorenzo triumph 23 May 10 | M
- Rossi injured in 16 Apr 10 | Mo




Home > Racing > MotoGP > News

**MotoGP**

15 Aug 2010

VALENTINO ROSSI AND DUCATI TOGETHER FROM 2011



Ducati and Valentino Rossi have signed a two-year agreement for the nine-times world Champion to race with the "Rossi" of Borgo Panigale, the Ducati MotoGP Team from 2011.



Names to be clustered could be represented with the Named Entities they co-occur in the same document.



# A general Approach

- An algorithm for clustering (such as Hierarchical Agglomerative Clustering)
- A fixed threshold determines how close two elements (i.e. documents or clusters) have to be so as to be grouped together.

e.g.:

```
cluster (d1, d2, threshold)
1  sim ← word_overlap(d1, d2)
2  if sim >= threshold
3      then return group(d1, d2)
4  return NIL
```



# A difficult Problem

- Discontinuity

A person may be mentioned in very different contexts and different time periods. e.g.

Montezemolo:	President of Ferrari
	President of Italia Futura
	Ex-President of Confindustria
	vice president of Bologna FC

- Clustering algorithms

For clustering methods (e.g. Agglomerative) the stop conditions are not known.



# Example 1

2 different contexts -> one person Montezemolo only

TUTTI PER LA «CONCERTAZIONE»

## Legge Biagi, Unione divisa

«Sono molto preoccupato per il Paese».

Nonostante i «modesti segnali di ripresa», per **Luca Cordero di Montezemolo** la tendenza positiva può diventare «effimera» se non è accompagnata da «scelte condivise, ma rigorose, coraggiose e, se serve, anche impopolari», a partire da un «drastico taglio» della spesa pubblica.

Perché, vista la situazione dei conti e dell'economia, la posta in palio è alta:

«L'**Italia** rischia di non farcela».

Quest'ultimo è il passaggio più allarmato della relazione con cui il numero uno di **Confindustria** ha aperto ieri a **Roma** l'annuale Assemblea pubblica degli industriali, spronando il nuovo Governo a imprimere quella svolta che per

## Massa confermato, Raikkonen no

Il presidente della **Ferrari Montezemolo**, alla vigilia di **Singapore**, dà fiducia al brasiliano e lascierebbe aperta la porta ad **Alonso**. La **Ferrari del 2010** ballerà ancora a ritmo di samba, ma su chi farà coppia con **Felipe Massa** non è dato sapere.

Certo è il volante di **Kimi Raikkonen** comincia a scricchiolare.

Nella settimana che porta dritto al gran premio di **Singapore**, il secondo nella storia del mondiale di **Formula 1** dopo l'esordio in notturna nella passata stagione, è il patron del Cavallino **Luca Cordero di Montezemolo** a dare corpo e sostanza ai tanti rumors sul mercato piloti della **Rossa**.

«Avremo una guida brasiliana che merita un'altra chance, visto che sta bene, sul resto stiamo riflettendo sulla scelta migliore, ma abbiamo ancora tempo.

Decideremo entro poche settimane»:

le parole del presidente nel corso di un intervento all'**Istituto di scienze militari e aeronautiche** di **Firenze**.

Si tratta di fatto della riconferma a pieno titolo di **Massa**, fermato dal





# Example 2

2 similar contexts -> 2 persons Paolo Rossi

IL SÜDTIROL PAREGGIA A FERRARA

La Spal rimonta due gol

Muove la classifica, stricandosi ancora in zona play-out con un 2 a 2 in trasferta il Südtirol del debuttante mister Pellegrino, il siracusano subentrato in settimana a Bolzano all'esonerato Sebastiani, sul campo di una Spal che ha rimontato due gol.

Nel posticipo dell'11ª giornata di ritorno in 1ª Divisione per la matricola biancorossa del ds trentino Luca Piazzi la vittoria è sumata nella ripresa.

Nel primo tempo a segno il centravanti Marchi al 23' e poi Fischmaller al 52' con due conclusioni di notevole efficacia in area.

Nel secondo tempo, però, nel giro di sette minuti gli estensi di Remondina riuscivano a rimettere in piedi il match, prima accorciando le distanze al 27' con un rigore trasformato dal neo entrato Volpe, e poi acciuffando il pari al 34' con la sfortunata autorete di Nazari, ex del Mezzocorona, in tentativo di anticipare Paolo Rossi (un omonimo del grande Pablito).

In classifica la Spal sale a 36, il Südtirol a 29.

La classifica del girone A: Gubbio (\*) 57 punti;

internazionale».

Così Paolo Rossi, vincitore del più importante riconoscimento per un giocatore nel 1982 dopo la vittoria ai Mondiali di Spagna, ha commentato il probabile successo di Buffon.

«Speriamo che lo vinca, se lo merita senza dubbio.

Ed è l'unico italiano che forse può arrivare al successo perchè - commenta Rossi - a livello di singoli non ne avevamo altri che potevano ambire al trionfo».

Il popolare Pablito esclude che oltre al mondiale dietro la probabile vittoria ci sia anche la scelta di Buffon di disputare il campionato di serie B con la Juventus.

«Non credo che un giocatore venga ripagato perchè si sceglie di giocare in serie C, piuttosto che in A o in B. Penso al valore del giocatore ed è un premio che si assegna per meriti e anche per un certo tipo di comportamento».

Così si confermerebbe la tradizione che nell'anno del Mondiale a vincere sia un protagonista della manifestazione iridata.

Nonostante Buffon sia un portiere.

«Di solito vengono premiati gli attaccanti o in giocatori in grado di fare gol».

Qual è la sua classifica di Rossi per le prime tre posizioni?

«Primo Buffon, poi Henry e Kakà».

Anceletti ha detto che Kakà è il più forte giocatore del mondo.

Il brasiliano vincerà il premio nei prossimi anni?

«Penso proprio di sì - il pronostico di Paolo Rossi -



# A dynamic Threshold?

- According to Popescu and Magnini [3][4] the more the ambiguity of the name in the corpus the more the information you need to disambiguate it.

Our hypothesis:

- using different values of threshold for ambiguous names and non ambiguous names could improve clustering.

Issue: the ambiguity of names in the corpus is unknown!



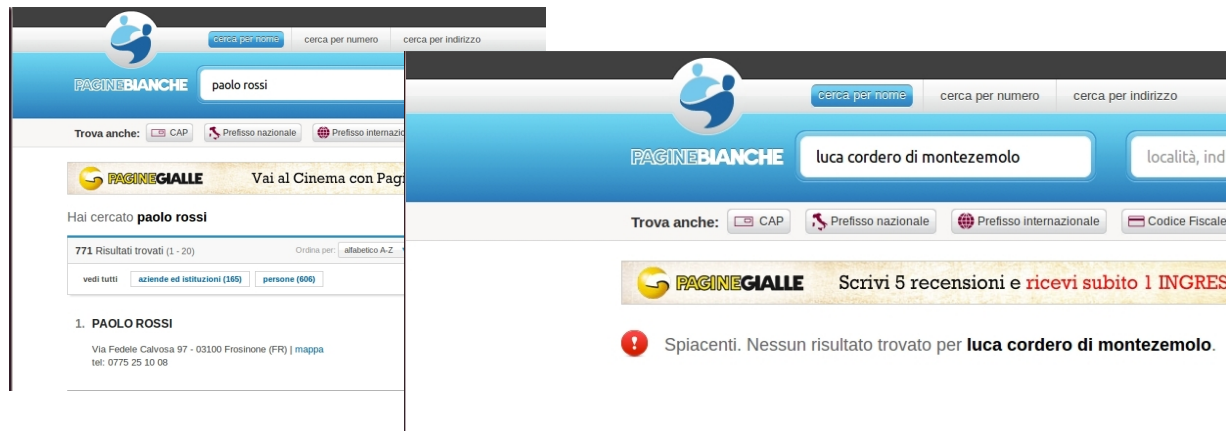
# NePS at EVALITA 2011

- Input: Set of documents matching a person name
- Output: Clusters, each cluster refers to the same individual
- System participants have to carry out the task for a number of unseen names
- System output is compared to gold-standard data



# Dynamic Threshold

As reported by Bentivogli et al. [1], PagineBianche (i.e. the Italian phonebook) is a good indicator of the ambiguity of a name in the NePS task: the more the ambiguity of a name in PagineBianche, the more the ambiguity of the name in NePS.



We could use PagineBianche to estimate the ambiguity of names.



# Dynamic Threshold

- Names categorized in 2 groups based on their ambiguity in PagineBianche:

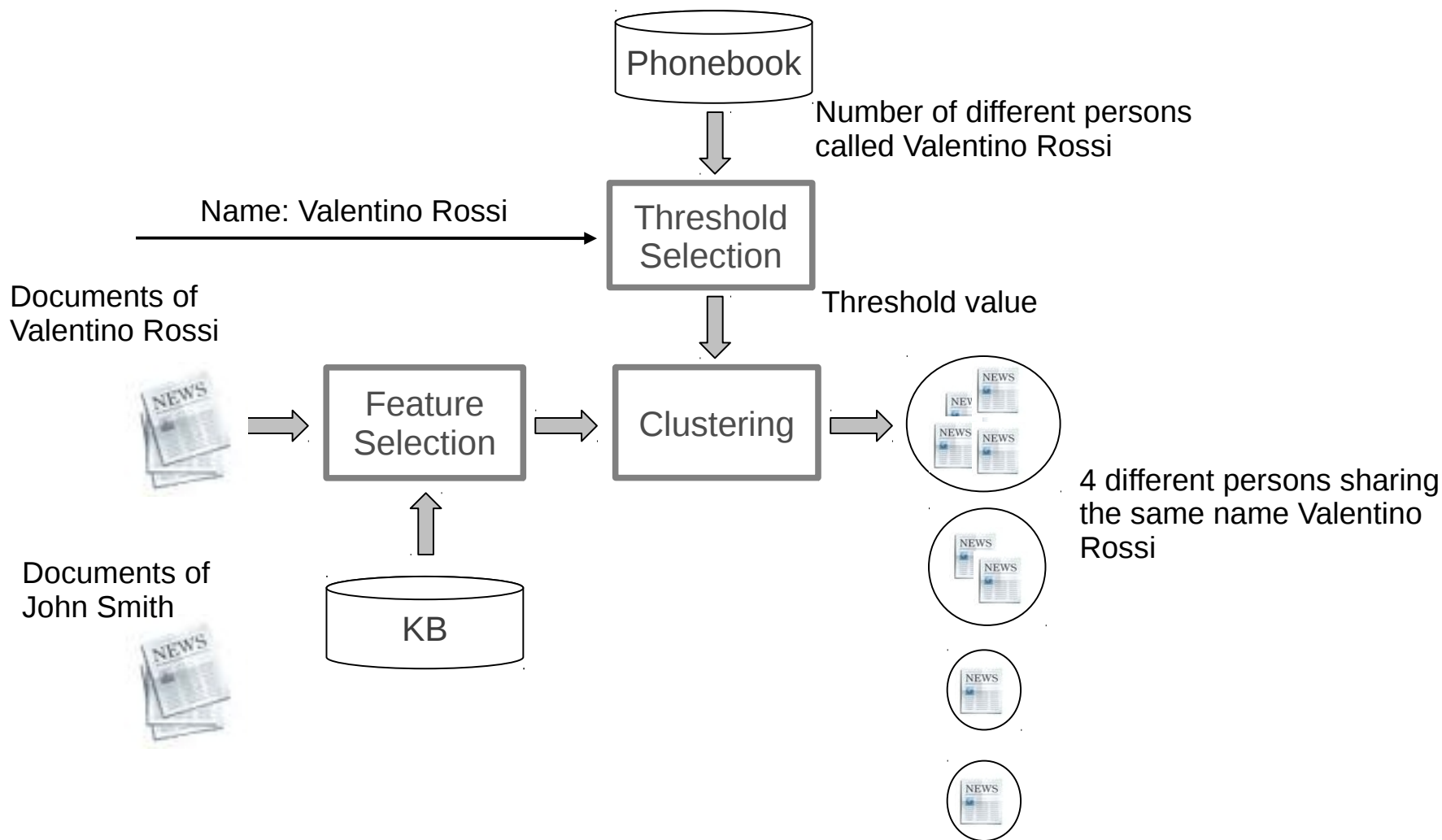
names occurring  $\leq 2$  -> non ambiguous names

name occurring  $> 2$  -> ambiguous names

- 2 values of threshold for ambiguous names and non ambiguous ones were calculated on the development set.
- threshold values were then used to cluster names of the test set.



# System Architecture





# Feature Set-1

- Topic of the Document: the topic the document is talking about (e.g. Political events, business, sports).
- Keyphrases: expressions, either single words or phrases, describing the most important concepts of a document (e.g. administrative committee, reduction in tax).
- Professional Category: professional category which is thought to belong to the name (e.g. president, journalist, football player).
- Named Entities: person, organization and location names (e.g. Bill Clinton, Ferrari, Rome).

Features extracted with  
TextPro ([textpro.fbk.eu](http://textpro.fbk.eu))



# Feature Set-2

Automatically linking person names in an ontology describing 30493 persons relevant to the Trentino region and national-level as well [5].

TUTTI PER LA «CONCERTAZIONE»

## Legge Biagi, Unione divisa

«Sono molto preoccupato per il Paese». Nonostante i «modesti segnali di ripresa», per **Luca Cordero** di **Montezemolo** diventare «effimero», le dimissioni di **Luca Cordero** di **Montezemolo** condivise, ma i «modesti segnali di ripresa», per **Luca Cordero** di **Montezemolo** impopolari», a pubblica. Perché, vista la sua in palio è alta: «L'Italia rischia di perdere il suo prestigio». Quest'ultimo è il cui il numero un **Luca Cordero** di **Montezemolo** l'annuale Assem nuovo Governo

## Massa confermato, Raikkonen no

Il presidente della **Ferrari Montezemolo**, alla vigilia di **Singapore**, dà fiducia al brasiliano e lascierebbe aperta la porta ad **Alonso**. La **Ferrari del 2010** ballerà ancora a ritmo di samba, ma su chi farà coppia con **Felipe Massa** non è dato sapere. Certo è il volante di **Kimi Raikkonen** comincia a scricchiolare. **Nella settimana** che porta dritto al gran premio di **Singapore**, il secondo nella storia del mondiale di **Formula 1** dopo l'esordio in notturna nella passata stagione, è il patron del Cavallino **Luca Cordero** di **Montezemolo** a dare corpo e sostanza ai tanti rumors sul mercato piloti della **Rossa**. «Avremo una guida brasiliana che merita un'altra chance, visto che sta bene, sul resto stiamo riflettendo sulla scelta migliore, ma abbiamo ancora tempo. Decideremo entro **poche settimane**». le **parole del presidente** nel corso di un intervento all'**Istituto di scienze militari e aeronautiche di Firenze**. Si tratta di fatto della riconferma a pieno titolo di **Massa**, fermato dal

## Background Knowledge

Nome: Luca  
Cognome: Cordero di Montezemolo  
Nato a: Bologna  
Nato il: 31/08/1947  
Titolo di studio: Laurea in Giurisprudenza  
Posizione attuale: Presidente della Ferrari

dal: 2004 al: 2010

Posizione: Presidente di Fiat S.p.a

dal: 2004 al: 2008

Posizione: Presidente di Confindustria

Linking accuracy; 71.50%  
Coverage: 21.35%





# Feature Weighting

A common choice is Inverse Document Frequency (IDF) where one intuition is at play: the more documents a feature  $f_k$  occurs in, the smaller its contribution is in characterizing the semantics of a document in which it occurs.



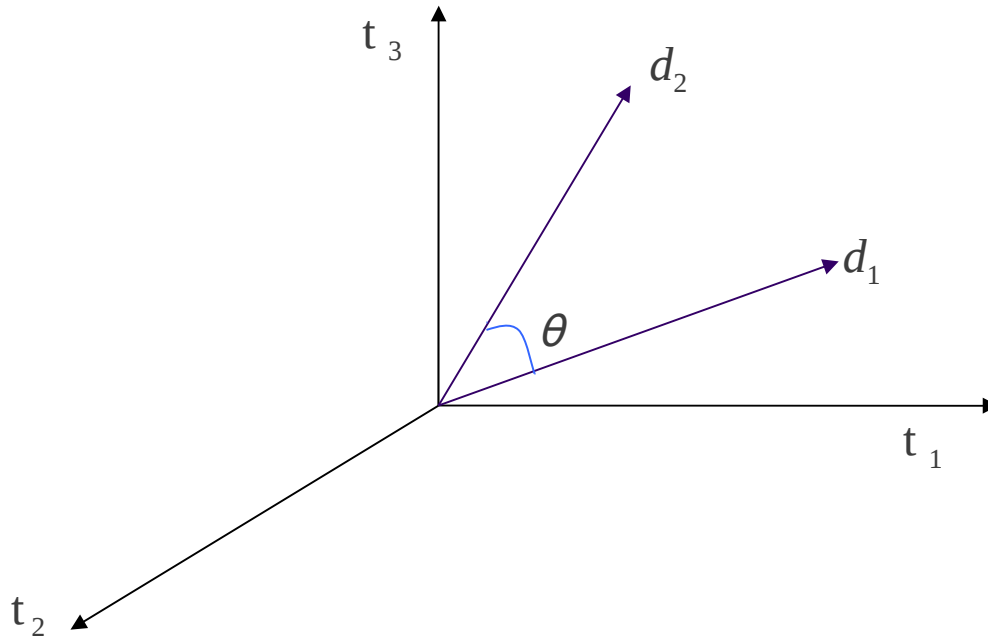
# QT

- Invented for gene clustering [2]
  - deterministic (differently from Hierarchical Agglomerative clustering)
  - does not require specifying the number of clusters a priori (as K-means does).
  - requires the a priori specification of the diameter (the distance between each pair of elements).
  - computationally Intensive,  $O(n^3)$  and Time Consuming



# Similarity Measure

- Distance between vectors  $d_1$  and  $d_2$  captured by the cosine of the angle  $x$  between them.





# Results

	All			No ambiguity			Medium ambiguity			High ambiguity		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
FBK	0.89	0.97	<b>0.93</b>	1.00	0.99	0.99	0.89	0.95	0.92	0.71	0.96	<b>0.82</b>
All-In-One	0.84	1.00	0.91	1.00	1.00	1.00	0.86	1.00	0.93	0.56	1.00	0.72

Bcubed precision, recall, and F1 measure for different levels of ambiguity of a name: no ambiguity, medium ambiguity and high ambiguity.



# T2 system/Applications



Based on QT

Dynamic Threshold

Implemented in java

Multithread

Accuracy: 93%<sup>(1)</sup>

Annotation speed:  
2.5M names per hour<sup>(2)</sup>

Available autumn 2012 as  
additional module of TextPro

(1) Evalita 2011

(2) PC Linux REDHAT 32GB RAM, 8  
cores

**L'Edicola della Conoscenza**  
Giovedì 24 Novembre 2011

**Text tagging**

VAL DI PEIO

Si **Alonso** in vetta con la **Ferrari** regina «bagnata»  
 Una «fantastica» giornata per riportare la **Ferrari** di **Fernando Alonso**  
 in testa al Mondiale a due ore dallo start.  
 Il primo in lancia per il **Boxer** il primo giro: **Checco del Nord**  
 Da con la quinta vittoria della stagione per **Maranello** che ottiene da  
**Luca Cordero di Montezemolo**.  
 Per lo spagnolo è il successo numero 26 in 57 gare che riempie di gioia il  
 pilota che in un giro di tempo di poche ore ha dimostrato che con la  
 determinazione, l'impegno, l'abilità e la voglia di vincere si riesce ad  
 uscire dalle situazioni più difficili.  
 Lui siamo una squadra, sostiene **Montezemolo**: «Non è mai o lo ha  
 fatto vedere ancora sulla volta».  
 «Il mio desiderio rimane con i piedi per terra».  
 Il compagno rimane aperto e accoglie il mondo degli  
 sponsor della gara.  
 Dovranno affrontare le ultime due gare con ancora più concentrazione ed

**Image tagging**

**Linking**

**Background Knowledge**

Nome: Luca  
 Cognome: Cordero di Montezemolo  
 Nato a: Bologna  
 Nato il: 23/09/1947  
 Titolo di studio: Laurea in Giurisprudenza  
 Posizione attuale: Presidente della Ferrari  
 dal: 2004 al: 2010  
 Posizione: Presidente di Fiat S.p.a  
 dal: 2004 al: 2005  
 Posizione: Presidente di Confindustria

733738	340147
33403	52478
2455	16649
1402	

Trentino  
Knowledgestore

T2 system was used in LiveMemories  
project (<http://www.livememories.org>)



# References

1. Bentivogli, L., Girardi, C., Pianta, E.: Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News. In: LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management (2008)
2. Heyer, L.J. and Kruglyak, S. and Yooseph, S.: Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Research*. 9, 1106-1115 (1999)
3. Octavian Popescu, Bernardo Magnini: Alleviating the Problem of Wrong Coreferences in Web Person Search. *CICLing* 280-293 (2009)
4. Octavian Popescu: Name Perplexity. *HLT-NAACL (Short Papers)* 153-156 (2009)
5. Tamilin, A., Magnini, B., Serafini, L.: Leveraging Entity Linking by Contextualized Background Knowledge: A case study for news domain in Italian. In: 6th Workshop on Semantic Web Applications and Perspectives (SWAP10), Bressanone, Italy (2010)



# Dev set results

NePS dev set:

T2: 0.937 (F1)

HAC, single-link: 0.750 (F1)

WePS-2 test set

Best system: 0.82(F1)

ALL-IN-ONE: 0.53

ONE-IN-ONE: 0.34



# QT

## Algorithm:

1. A random gene is chosen from the selected gene list.
2. The algorithm determines which gene has the greatest similarity to this gene. If their total diameter does not exceed the diameter threshold, then these two genes are clustered together.
3. Other genes that minimize the increase in cluster diameter are iteratively added to this cluster. This process continues until no gene can be added to this first candidate cluster without surpassing the diameter threshold.
4. A second candidate gene is chosen.
5. The algorithm determines which gene has the greatest similarity to this second gene. **All genes in the selected gene list are available for consideration to the second candidate cluster.**
6. Other genes from the selected gene list that minimize the increase in cluster diameter are iteratively added to the second candidate cluster. The process continues until no gene can be added to this second candidate cluster without surpassing the diameter threshold.
7. The algorithm iterates through all genes on the selected gene list and forms a candidate cluster with reference to each gene. In other words, there will be as many candidate clusters as there are genes in the gene list. Once a candidate cluster is formed for each gene, all candidate clusters below the user-specified minimum size are removed from consideration.
8. The largest remaining candidate cluster, with the user-specified minimal number of gene member, is selected and retained as a QT cluster. The genes within this cluster are now removed from consideration. All remaining genes will be used for the next round of QT cluster formation.
9. The entire process (step 1 to 9) is repeated until the largest remaining candidate cluster has fewer than the user-specified number of genes.