Named Entity Recognition on Transcription using cascaded classifiers

Firoj Alam[1,2]

[1]SIS Lab, Department of Information Engineering and Computer Science, University of Trento

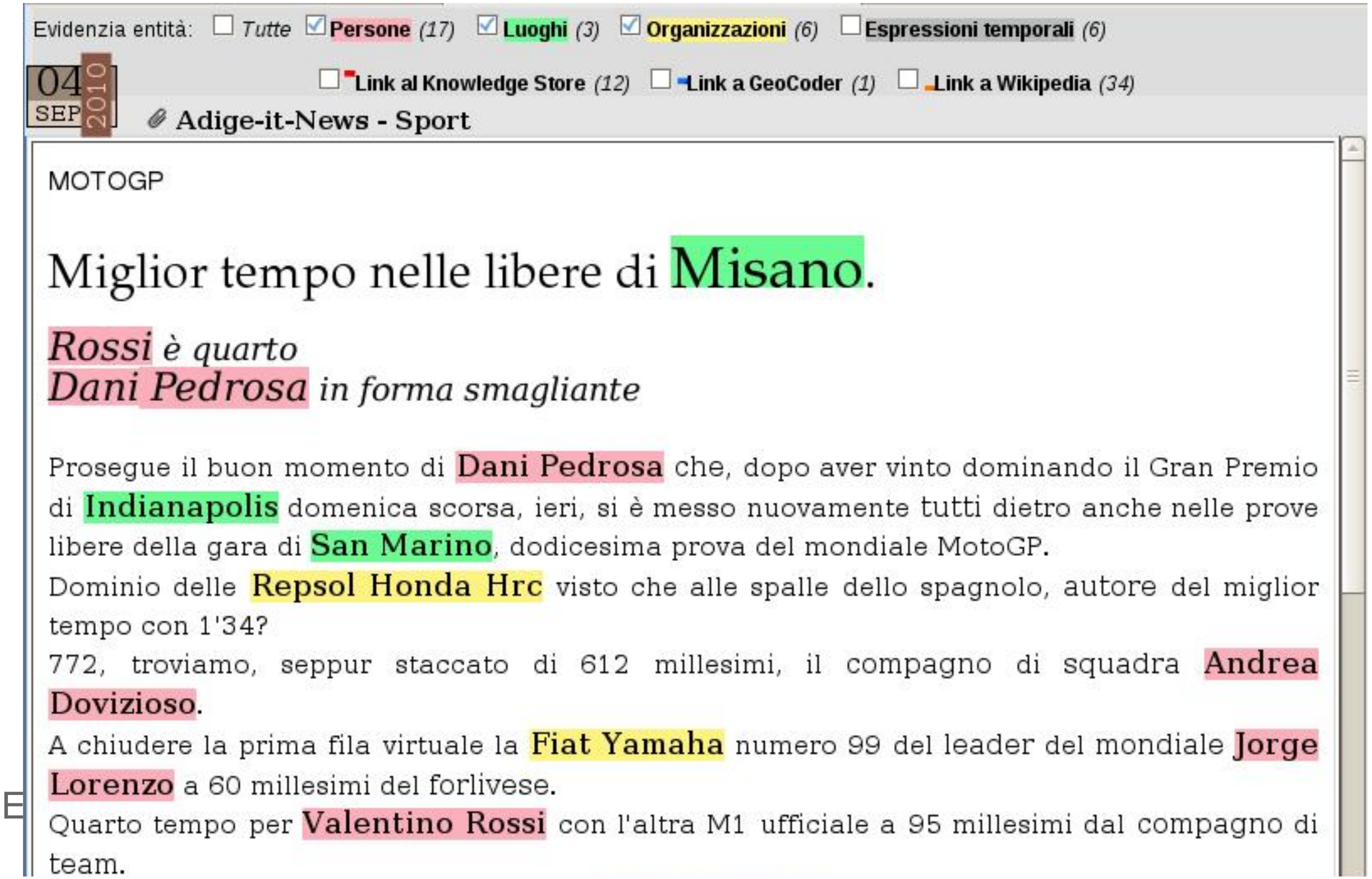
[2]HLT Unit, FBK-irst

Overview

- Named Entity Recognition (NER)
- Experiment
  - Written vs spoken documents (e.g. transcription)
  - System architecture
  - Case restoration model
- Results
- Conclusion and future study

Named Entity Recognition system

NER is the subtask of Information Extraction (IE) aiming to detect and classify entities in texts into predefined categories such as person, location, organization, time expressions and so on.



Evidenzia entità: □ *Tutte* ☑ **Persone** *(17)* ☑ **Luoghi** *(3)* ☑ **Organizzazioni** *(6)* □ **Espressioni temporali** *(6)*

□ **Link al Knowledge Store** *(12)* □ **Link a GeoCoder** *(1)* □ **Link a Wikipedia** *(34)*

04 SEP 2010

Adige-it-News - Sport

MOTOGP

Miglior tempo nelle libere di Misano.

Rossi è quarto
Dani Pedrosa in forma smagliante

Prosegue il buon momento di Dani Pedrosa che, dopo aver vinto dominando il Gran Premio di Indianapolis domenica scorsa, ieri, si è messo nuovamente tutti dietro anche nelle prove libere della gara di San Marino, dodicesima prova del mondiale MotoGP.
Dominio delle Repsol Honda Hrc visto che alle spalle dello spagnolo, autore del miglior tempo con 1'34?
772, troviamo, seppur staccato di 612 millesimi, il compagno di squadra Andrea Dovizioso.
A chiudere la prima fila virtuale la Fiat Yamaha numero 99 del leader del mondiale Jorge Lorenzo a 60 millesimi del forlivese.
Quarto tempo per Valentino Rossi con l'altra M1 ufficiale a 95 millesimi dal compagno di team.

Written vs Spoken documents

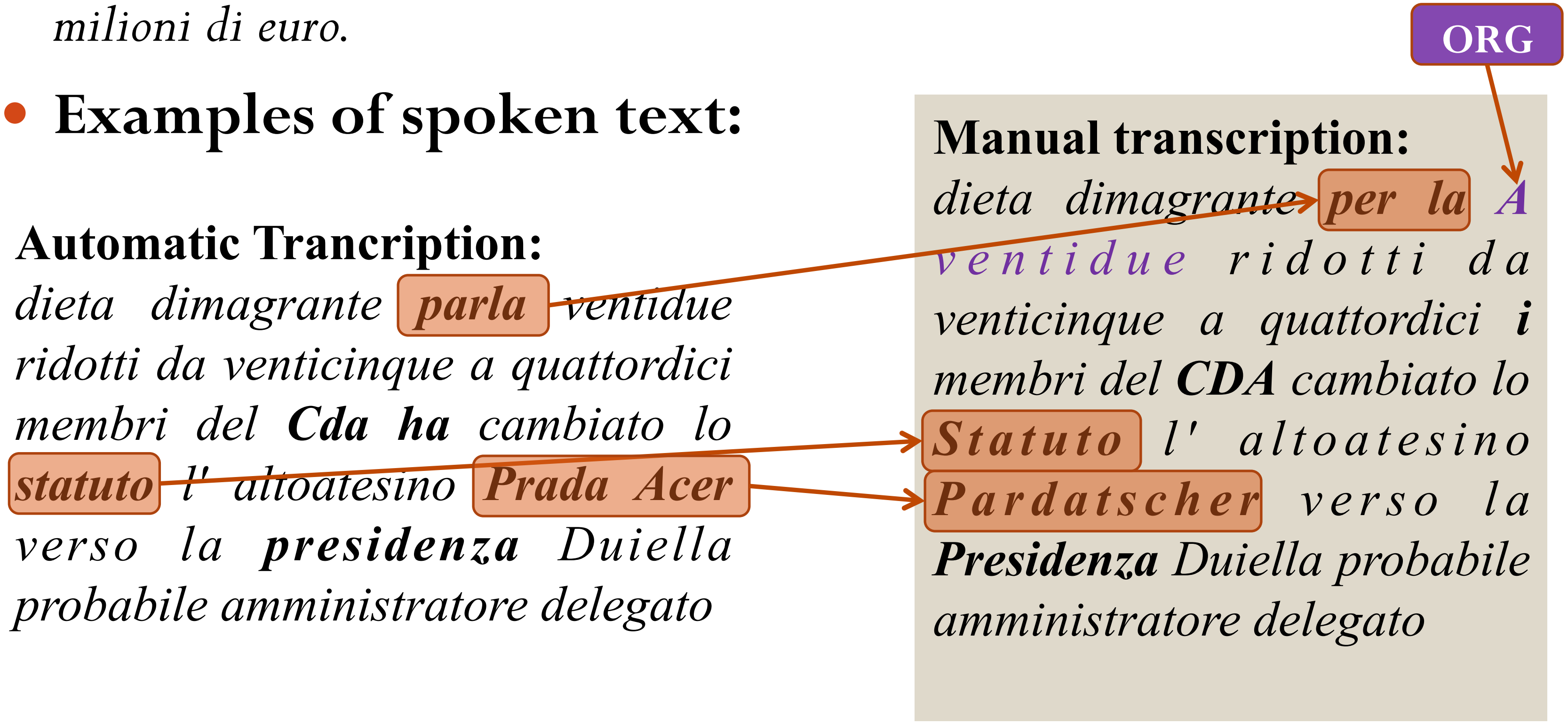
- **Written documents:** Text appears as standard written form e.g. newspaper articles.
- **Spoken documents:** Speech (e.g. broadcast news) are transcribed using Automatic Speech Recognition (ASR) system.
- Three factors of recognizing NEs in spoken documents:
  - Case information is missing
  - Punctuation marks is missing
  - ASR errors

Written vs Spoken documents

- **Examples of written text:**

Dal 2000 ad oggi sono stati così sottratti alle casse dello Stato ben 14 milioni di euro.

- **Examples of spoken text:**



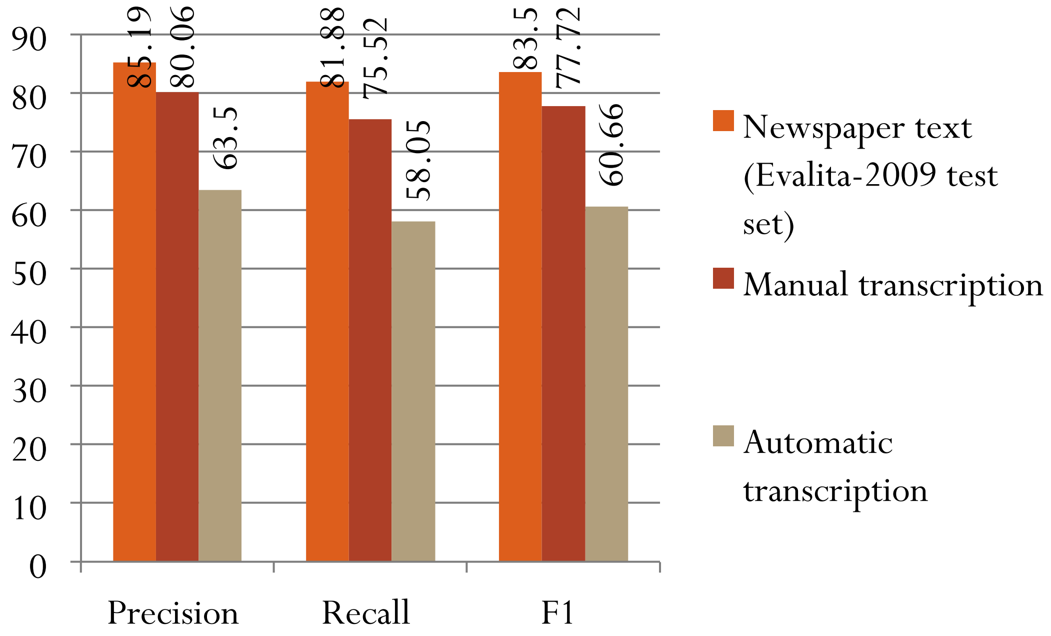
Automatic Trancription:
*dieta dimagrante **parla** ventidue ridotti da venticinque a quattordici membri del **Cda ha** cambiato lo **statuto** l' altoatesino **Prada Acer** verso la **presidenza** Duiella probabile amministratore delegato*

ORG

Manual transcription:
*dieta dimagrante **per la** A ventidue ridotti da venticinque a quattordici **i** membri del **CDA** cambiato lo **Statuto** l' altoatesino **Pardatscher** verso la **Presidenza** Duiella probabile amministratore delegato*

Written vs Spoken documents

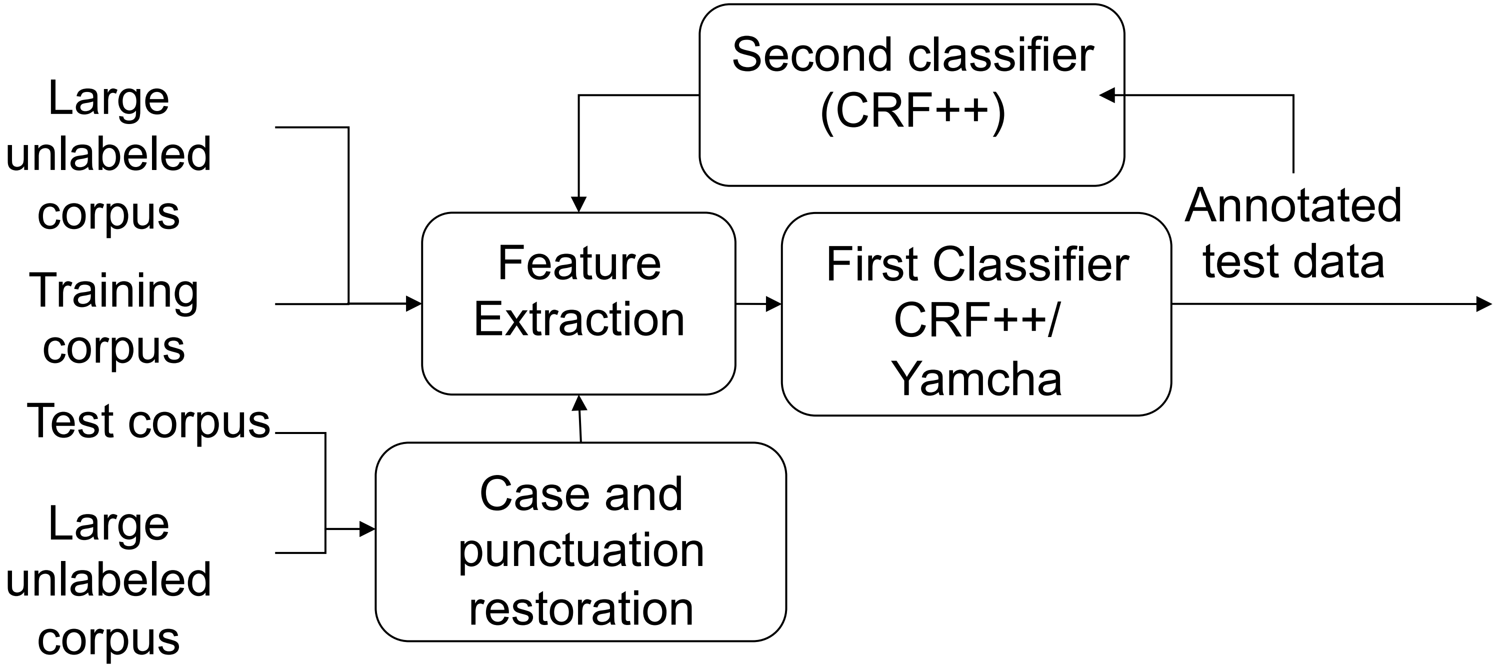
- Final classification has been done using Yamcha



- The word error rate (WER) of the ASR is **16.39%,** unit accuracy is **83.61%** and percent correct is **87.48%**

System Architecture

- Approach is similar to Typhoon developed by HLT unit at FBK.
- Second classifier is based on CRF instead of HMM



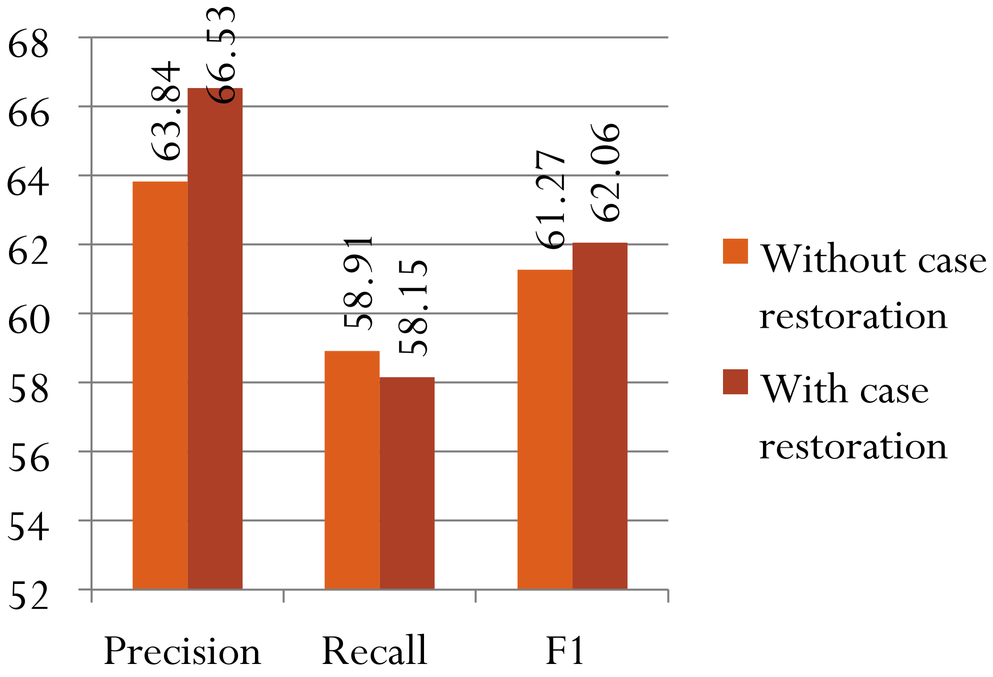
Second Classifier

- Using unlabeled datasets as additional features

  1. *First classifier (CRF++)* is trained on annotated data (training set)
  2. Annotate unlabeled data by *first classifier*
  3. *Second classifier* is trained on datasets that is produced by first classifier in step 2 and it classifies training and test sets to integrate additional features.
  4. Finally, retrain the *first classifier* on the training set produced in step 3 and classify the test data

Case and Punctuation Restoration

- L'adige corpus
- Classifier is based on CRF
- Performance of this model is 96.49



Official results on closed task

Category	**Precision**	**Recall**	**F1**
Overall	61.76%	60.23%	60.98
GPE	81.79%	78.52%	80.12
LOC	65.22%	47.87%	55.21
ORG	50.21%	43.85%	46.82
PER	47.28%	55.26%	50.96

Official results on Open task

Category	**Precision**	**Recall**	**F1**
Overall	65.55%	61.69%	63.56
GPE	80.33%	80.44%	80.38
LOC	76.36%	44.68%	56.38
ORG	60.51%	47.52%	53.24
PER	48.92%	54.39%	51.51

Official results of manually transcribed test set

Task	**Precision**	**Recall**	**F1**
Closed task	79.33%	79.80%	79.57
Open task	82.82%	81.27%	82.04

Conclusion and future study

- Case and punctuation model improve the performance of the system
- Exploiting unlabeled datasets helps to improve the performance
- Future Study:
  - Using unlabeled transcribed data
  - Adapting relevant sentences from unlabeled data
  - This system is going to include into typhoon which is available as a part of Textpro **(http://textpro.fbk.eu/**).

Thank you

???