Creating a Coreference Resolution System for Italian

Massimo Poesio, Olga Uryupina
CIMEC, University of Trento

Yannick Versley
University of Tuebingen

19.05.2010

Outline

- Coreference Resolution
- BART
- Adapting BART to Italian
- Evaluation experiments: Evalita-2009

Coreference Resolution

One reason **Lockheed Martin Corp.** did not announce a full acquisition of Loral Corp. on Monday, according to Bernard Schwartz, Loral's chairman, was that **Lockheed** could not meet the price he had placed on Loral's 31 percent ownership of Globalstar Telecommunications Ltd.

Coreference Resolution

One reason Lockheed Martin Corp. did not announce a full acquisition of **Loral Corp.** on Monday, according to Bernard Schwartz, **Loral**'s chairman, was that Lockheed could not meet the price he had placed on **Loral**'s 31 percent ownership of Globalstar Telecommunications Ltd.

Coreference Resolution

One reason Lockheed Martin Corp. did not announce a full acquisition of Loral Corp. on Monday, according to **Bernard Schwartz**, Loral's **chairman**, was that Lockheed could not meet the price **he** had placed on Loral's 31 percent ownership of Globalstar Telecommunications Ltd.

Coreference Resolution

Goal: identify “entities” or “chains”

- {Lockheed Martin Corp., Lockheed}
- {Loral Corp., Loral, Loral}
- {Bernard Schwartz, chairman, he}
- {Monday}
- ..

Input: raw text + set of “mentions” (“markables”)

Why do we need it?

- Information Extraction
- IR/QA
- Machine Translation
- Summarization
- Reverse task – generating anaphoric expression, e.g. to improve coherence in a generated document

Prerequisites

- Set of mentions – Mention Detector needed
- Linguistic Information – different layers of lingustic knowledge (PoS, morphology, parse trees, semantic labels,..)

Multilinguality

- Preprocessing – varying quality
- Till now – only very few non-English systems
- SemEval-2010

Our approach: take a modular system (BART) and extend it to cover another language

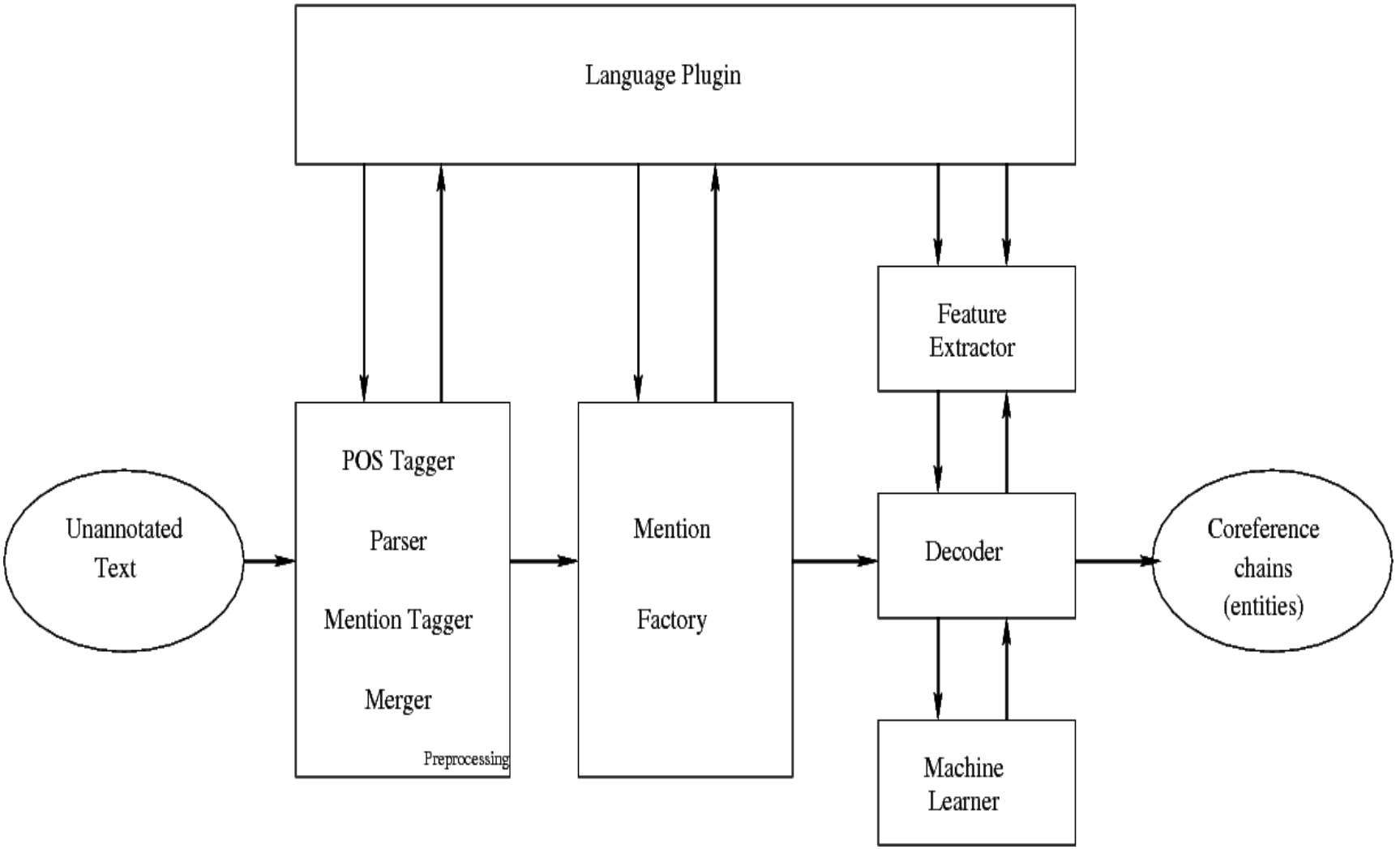
BART

Baltimore Anaphora Resolution Toolkit

- Prototype: Ponzetto & Strube (2006)
- 1st version: John Hopkins Workshop (2007)
- Current state: multiple features, several models, 3 languages supported, different input/output formats, 2 scoring metrics
- Evalita-2009 – the only system able to perform CR in Italian
- SemEval-2010 – state-of-the-art performance for En, It; best for De

BART's strong sides

- Modularity
- LanguagePlugin
- Extensive scoring/testing facilities
- Supports various input formats (including raw text – can be run from scratch)
- A lot of solutions already implemented


Language Plugin
Feature Extractor
Unannotated Text
POS Tagger
Parser
Mention Tagger
Merger
Preprocessing
Mention Factory
Decoder
Coreference chains (entities)
Machine Learner

Language-specific issues: Aliasing

- New aliasing patterns for PERSON/ORG:

  S.p.a., D.ssa

- Mining aliasing patterns for LOCATION/GPE:

  Verona → Citta di Verona

- Revising abbreviation constraints

Aliasing: evaluation

- “Universal” aliasing

  MUC: R=17.2, P=79.2, F=28.3
- “Italian” aliasing

  MUC: R=22.5, P=90.7, F=36.0

Language-specific issues: preprocessing

- Parsing vs. Shallow pipeline
- EMD – Biggio et al. 2009
- Manipulating training data to reduce noise – adjusting mention boundaries, removing too complex NPs

Preprocessing: evaluation

- Parsing pipeline (+ SemClass filtering)

  MUC R=42.4, P=73.7, F=53.8
- Shallow pipeline

  MUC R=45.8, P=72.3, F=56.1

  Parsing Pipeline no reliable – not enough training material

Evalita-2009

- BART – the only system able to perform the task

	R	P	F
MUC	45.8	72.3	56.1
CEAF	62.1	64.6	63.6

Evalita

- Language-agnostic vs. Italian settings (gold mentions, MUC)

	R	P	F
Universal	34.9	76.6	47.9
Italian	46.8	71.1	56.4

Conclusion

- Extending BART to cover Italian
- Re-training (“universal”) possible, but..
- Better results by addressing language-specific issues:
  - language-specific aliasing
  - shallow preprocessing
- Possible because of the modular design
- Mid-scale project, other languages are waiting for

Thank you!