



EVALITA 2011

Evaluation of NLP and Speech Tools for Italian

EVALITA 2011

Anaphora Resolution

Olga Uryupina, Massimo Poesio



UNIVERSITÀ DEGLI STUDI
DI TRENTO



LiveMemories
Active Digital Memories of Collective Life



University of Essex



Outline

- Introduction
- LM-Wikipedia dataset
 - Annotation guidelines
 - Corpus statistics
- Evaluation
 - Metrics
 - System results
- Conclusion and future work



Intro: Anaphora Resolution

- Task: identify and link together all the mentions of the same real-world entity in a document

Enrico Conci (Trento, 24 giugno 1866 – Trento, 25 marzo 1960) è stato un politico italiano. Ha sostenuto l'autonomia trentina quando la provincia era ancora sotto l'Impero Austro-Ungarico ed è poi stato senatore del Regno d'Italia e della Repubblica Italiana.



Intro: Anaphora Resolution

- Task: identify and link together all the mentions of the same real-world entity in a document

Enrico Conci (Trento, 24 giugno 1866 – Trento, 25 marzo 1960) è stato un politico italiano. Ha sostenuto l'autonomia trentina quando la provincia era ancora sotto l'Impero Austro-Ungarico ed è poi stato senatore del Regno d'Italia e della Repubblica Italiana.



Intro: Anaphora Resolution

- Task: identify and link together all the mentions of the same real-world entity in a document

Enrico Conci (Trento, 24 giugno 1866 – Trento, 25 marzo 1960) è stato un politico italiano. **Ha** sostenuto l'autonomia trentina quando la provincia era ancora sotto l'Impero Austro-Ungarico ed **è** poi stato senatore del Regno d'Italia e della Repubblica Italiana.



Intro: Anaphora Resolution

- Task: identify and link together all the mentions of the same real-world entity in a document

Enrico Conci (Trento, 24 giugno 1866 – Trento, 25 marzo 1960) è stato un politico italiano. Ha sostenuto l'autonomia trentina quando la provincia era ancora sotto l'Impero Austro-Ungarico ed è poi stato senatore del Regno d'Italia e della Repubblica Italiana.



Intro: Anaphora Resolution

- Prerequisite for higher-level tasks:
 - Summarization
 - Information Extraction
 - Question Answering
 - ..
- Complex task on its own, relies on other NLP components:
 - NER
 - Parsing
 - Encyclopedic knowledge
 - ..



Intro: state of the art

- English
 - Many corpora, different annotation guidelines
 - ACE
 - OntoNotes
 - Many systems, models,..
 - Features
- Italian
 - LEDR task (Evalita-2009)
 - SemEval-2010
 - Very few systems
 - Preprocessing difficulties



Intro: AR vs LEDR

- LEDR (Evalita 2009)
 - ACE-style annotation
 - Restricted set of semantic types
 - Only referential NPs
- Anaphora Resolution (Evalita 2011)
 - OntoNotes-style annotation
 - All nominal expressions
 - Non-referential NPs covered
 - More linguistic phenomena

=> More complex task



LM-Wikipedia corpus

- Annotated within the LiveMemories project (UniTN+FBK+..)
- Wikipedia articles (Trentino)
- Following ARRAU/OntoNotes annotation style
- Guidelines adjusted for Italian (zero pronouns, citics, etc)
- Formats:
 - MMAX XML
 - CoNLL/SemEval tabular
- Competitions: SemEval-2010, Evalita-2011
- Rodriguez et al (2010)



LM-Wikipedia guidelines

Cover all nominal mentions (“markables”)

- Referring
 - New
 - Anaphoric
- Non-referring
 - Expletive
 - Idiom
 - Predicative
 - Coordination



LM-Wikipedia guidelines

Cover all semantic types

- Abstract
- Person
- None (non-referring)
- Concrete
- GSP
- Time
- Location
- Organization
- Facility
- Numeric
- Animate
- unknown



LM-Wikipedia: issues

Predicative NPs

- Predication

Enrico Conci e' stato un politico italiano.

- Apposition

Enrico Conci, un politico italiano, ..

Different annotation styles

- ACE: treat as coreference
- OntoNotes: ignore
- ARRAU: treat as non-referring



LM-Wikipedia: issues

Zero pronouns

Enrico Conci (Trento, 24 giugno 1866 – Trento, 25 marzo 1960) è stato un politico italiano. **Ha** sostenuto..



LM-Wikipedia: issues

Coordination

Giovanni e Giuseppe si incontrarono. I due ragazzi andarono al cinema.

- MMAX: “i due ragazzi” → {Giovanni, Giuseppe}
- CoNLL: “i due ragazzi” no antecedent



LM-Wikipedia: issues

Discontinuous markables

Enrico ed Elsa Conci

3 markables: “Enrico ed Elsa Conci”, “Enrico .. Conci”,
“Elsa Conci”

studenti e docenti dell' Università di Trento



LM-Wikipedia: statistics

- 3 datasets: training (100k tokens, 30k markables), SemEval-test (40k), Evalita-test (25k)
- 4th dataset being annotated now (35k)

- Referentiality

referential, new	17607
referential, anaphoric	8422
non-referential	4279



LM-Wikipedia: statistics

- Nominal vs. verbal markables

nominal	29047
---------	-------

verbal	1262
--------	------

- Discontinuous markables 306



LM-Wikipedia: statistics

- Semantic types

abstract	7211
person	5072
- (non-ref)	4263
concrete	2798
gsp	2710
...	

! ACE types are too restrictive



Evaluation: metrics

- No agreement on the best metric in the field
- CoNLL/SemEval metrics: MUC, B3, CEAF_e, CEAF_m, BLANC
- CoNLL/SemEval scorer
- Adjustment: more relaxed mention alignment (ACE style)



Evaluation: results

- 1 participant (Uni Pisa), 2 runs

	MUC	B3	CEAFm	CEAF _e	BLANC
Run 1	26.36	83.79	72.99	78.89	55.94
Run 2	25.07	83.64	72.53	78.38	55.80



Conclusion

- Anaphora resolution track
 - New dataset (publicly available)
 - Cover wide variety of linguistic phenomena
 - CoNLL/SemEval format: easy to test on, no adjustments needed
- Very complex task
 - Just one participant submitted the run



Thanks!

We would like to thank LiveMemories, iKernels and PARLI for supporting this research