EVALITA 2011

Evaluation of NLP and Speech Tools for Italian

EVALITA 2011

UNIPI participation to the Anaphora Resolution Task

G. Attardi, S. Dei Rossi, M. Simi

Dipartimento di Informatica, Università di Pisa

Background and motivation

- Participation in “Coreference Resolution in Multiple Languages” task at SemEval-2010
  - Multiple languages: English, German, Spanish, Catalan, Italian …
  - *Gold standard* scenario: achieved top score for German
  - *Regular* scenario: achieved top score for Catalan and Spanish
- Motivation
  - Test system on the new Italian Corpus
  - Compare with other languages

Approach

- Differences and problems
  - No gold data provided for lemmas, PoS and parsing
  - Lack of Named Entities in the test set (key feature in Semeval)
  - No clear guidelines for determining mention boundaries
  - No official scorer provided, we used the Semeval scorer for tuning the system
- Split co-reference resolution into two sub-problems:
  1. Mention identification
  2. Clustering mentions referring to same entity

Mention identification: strategy

- Based on analysis of parse trees
  - Lack of reliable parse tree: retagging with Tanl suite
  - Lack of criteria for "*verbal*" one-token mentions (~1200)
  - Lack of correspondence of mentions with sub-trees of parse tree
- Dedicated classifier for verb mentions and dates
  - ME Tanl classifier with specific features
- Mention detection
  - verbs and dates identified by the ME classifier;
  - subtrees with heads: *common and proper nouns*; *personal*, *demonstrative*, *indefinite*, *possessive pronouns*
  - several heuristics for deciding what to include and what to exclude wrt to parse sub-tree …

Heuristics for mention boundaries

- Used in *Run1* and *Run2:*
  - \+ include articulated preposition at the beginning of mentions;
  - –exclude clitic pronouns at the beginning of mentions;
  - –stop right mention expansion on balanced punctuation and on commas when the parser relation is coordinate conjunction;
  - –remove articulated preposition and relative pronouns from the right boundary of mentions;
  - –remove preposition and balanced punctuation from the left boundary of mentions;
- Used only in *Run1*, in an attempt to improve precision:
  - –when dependency relation is “modifier”, consider as head of NPs only nouns and pronouns

Determining coreference

- Trained a binary ME classifier to decide whether two mentions refer to the same entity
  - Positive examples: any mention together with each preceding mention with the same number (referring to same entity)
  - Negative examples: any mention together with each preceding mention with different number
- Features
  - Lexical Features: *same, prefix, suffix, acronym, edit distance*
  - Distance Features: *sentence, token, mention distance*
  - Syntax Features*: same head PoS, pairs of head PoS*
  - Count Features: pairs of number of occurrences of mentions
  - Pronoun Features: *gender, number, pronoun type*

Mention clustering

- *Best-first greedy clustering* algorithm
  - Each mention is compared to all previous mentions (collected in a global mentions table)
  - If the pair-wise classifier assigns a probability greater than a given threshold when comparing with a previously identified entity, it is assigned to that entity.
  - In case more than one entity has a probability greater than the threshold, the mention is assigned to the one with highest probability.

Evalita official results

	Run 1			Run 2		
	Recall	**Precision**	**FB1**	**Recall**	**Precision**	**FB1**
Ident. of ment.	64.01%	62.11%	63.04	64.12%	59.36%	61.65
MUC	18.38 %	46.59%	26.36	17.83 %	42.21%	25.07
B-CUB	75.69%	93.83%	83.79	75.96%	93.04%	83.64
CEAFm	72.99%	72.99%	72.99	72.53%	72.53%	72.53
CEAFe	87.64%	71.72%	78.89	86.53%	71.64%	78.38
BLANC	53.75%	64.66%	55.94	53.66%	64.38%	55.80

Run 1 is the best run, more precise in mention identification

Discussion

	SemEval scorer		Evalita scorer	
	dev	test	dev	test
Identification of mentions	71.83	67.34	64.21	63.04
Coreference (B-CUB)	65.99	59.37	84.74	83.79

- Identification of mentions proved to be difficult :
  - most data were system predicted (not gold);
  - heuristics were not effective, due to our own poor understanding of annotation guidelines
  - in particular the model trained to recognize those verbs that are also mentions, effective on the dev set, failed badly to predict on the test set: 29% recall, 18% precision.

Conclusion

- Results cannot be compared with other participants
- Results cannot be compared with the results obtained in SemEval-2010
  - The task is different: lack of gold data, NE's, … more difficult and ill defined in some aspects
  - The scorer is different: apparently more strict in mention detection and more tolerant in coreference (it allows partial alignment between system and gold mentions).