The Berkeley Parser at the EVALITA 2011 Constituency Parsing Task

Alberto Lavelli

FBK-irst, Trento, Italy

24 January 2012

Participation in EVALITA constituency parsing task

Part of a wider research effort with Anna Corazza and Giorgio Satta devoted to

- application of state-of-the-art statistical parsing techniques to Italian (TLT 2004, EVALITA 2007 & 2009)
- exploration of information-theoretic measures that account for the empirical difference of the experimental results on different treebanks/languages (not yet published)

State-of-the-art Statistical Parsers (in 2004)

- Dan Bikel's parser – lexicalized
  - head-driven
  - splits RHS in *n* relations with the rule head
- Stanford parser (Klein & Manning) – unlexicalized
  - adds annotations to nodes to take context into account
  - rule Markovization to cope with data sparsity

Both parsers can be (and have already been) applied to different languages

Bikel/Stanford Parser – Adaptation to Italian

- identification of rules for finding lexical heads
- selection of a lower threshold for unknown words

No further language-dependent adaptations:

- for Bikel's parser, no tree transformations analogous to those introduced by Collins for the PennTreeBank
- for the Stanford parser, only basic annotations, i.e., parent annotation for both nonterminals and PoS tags and horizontal Markovization

Initial Results on Italian - TLT 2004

Application of state-of-the-art statistical parsing techniques to **Italian**

- **experimental results** with
  - different parsing methods
    - Bikel's parser
    - Stanford parser
  - on a single treebank
    - Italian Syntactic-Semantic Treebank (ISST, TLT 2004 paper)

	LR	LP	F_1
Bikel < 40	68.58	68.40	68.49
Stanford best < 40	66.31	62.19	64.18

Initial Results on Italian

Results on ISST much worse than on English (and also worse than on other languages, e.g. Chinese, Czech, German)
Two possible explanations for the gap in performance:

- intrinsic differences between the two languages
- differences between the annotation policies adopted in different treebanks

Two lines of research:

- same experiments on a different Italian treebank (TUT: Turin University Treebank)
- exploring information theoretic measures to compare the difficulty of different parsing tasks

Measuring Parsing Difficulty

- New measure, called Expected Conditional Cross-Entropy (ECC), for comparing parsing difficulty across treebanks
- Conjecture: ECC strictly related to parsing performance
- ECC as an effective measure of parsing difficulty
- Conjecture tested comparing ECC and standard performance measures (P/R/F_1/ExactMatchRate) on treebanks for English (WSJ), French (FTB), German (Negra, TüBa-D/Z) and Italian (ISST, TUT)

EVALITA 2007 – Results

	LR	LP	F_1	EMR
EVALITA 2007	70.81	63.35	67.96	
Bikel test	71.73	69.88	70.79	9.05
Bikel test < 40	72.04	70.08	71.05	9.84
Stanford best	61.19	62.25	61.72	5.00
Stanford best < 40	63.03	64.23	63.62	5.43
ISST				
Bikel < 40	68.58	68.40	68.49	
Stanford best < 40	66.31	62.19	64.18	

EVALITA 2009

Berkeley parser (Petrov & Klein)

- based on a hierarchical coarse-to-fine parsing, where a sequence of grammars is considered, each being the refinement, i.e. a partial splitting, of the preceding one.
- no need for language-specific adaptations
- state-of-the-art performance for English on the Penn Treebank
- outperforms other parsers on German and Chinese (Petrov & Klein NAACL 2007), and French (Seddah et al. IWPT 2009)

EVALITA 2009 – Results

	LR	LP	F_1	EMR
Bikel	68.51	64.45	66.42	14.00
Bikel < 40	68.99	65.03	66.95	14.81
Berkeley - iteration #4	**80.02**	**77.48**	**78.73**	**21.00**
Berkeley - iteration #4 < 40	79.90	77.92	78.90	22.22

EVALITA 2011

- Using again the Berkeley parser.
- Didn't manage to explore reranking and self-training to improve performance (lack of time)

EVALITA 2011 – Results

10-fold cross validation on the training set

	LR	LP	F_1	EMR
Berkeley	78.74	79.32	78.99	26.33
Berkeley < 40	81.88	82.38	82.10	31.82

Results on the test set

	LR	LP	F_1	EMR
Berkeley	83.54	84.12	83.83	22.74
Berkeley < 40	83.69	84.35	84.02	24.20

Conclusions and Future Work

- Improvement w.r.t. the previous editions of EVALITA
- Performance on Italian now at a reasonable level (given the limited size of TUT)
- Exploring reranking and the use of self-training to improve performance
- Berkeley parser ready for integration in the TextPro NLP suite (`http://textpro.fbk.eu`)

Thanks to Dan Bikel, Chris Manning and his colleagues, and Slav Petrov for making their parsers available.