Minimally Supervised Domain-Adaptive Parse Re-ranking for Relation Extraction

Feiyu Xu, Hong Li, Yi Zhang, Hans Uszkoreit and Sebastian Krause

DFKI, LT-Lab

Outline

- Motivation
- Ingredients
- Background
  - HPSG
  - DARE
- Parse Re-ranking
- Experiments
- Conclusion and future work

Motivation: Domain Adaptation

- Adaptation of a generic parser to a given relation extraction task or domain with minimal domain knowledge without actually changing the parser itself
- Constrution of a parse re-ranking model based on the confidence values of relation extraction rules automatically learned from the n-best parses
- Improving the parse selection with the parse re-ranking model, in order to obtain the best first parses for relation extraction task
- Evaluation of parse re-ranking concerning relation extraction and parsing

Ingredients

- Generic parser, grammar and treebank
  - ERG (Flickinger 2000)
  - PET parser (Calmeier, 2002)
  - Redwood treebank (Oepen et al., 2002)
- DARE: Framework for minimally supervised machine learning of relation extraction (RE) rules (http://dare.dfki.de)
  - Semantic seed as minimal domain knowledge
  - Each learned RE rule is assigned with confidence estimation
- Data for experiments and evaluation
  - DARE Nobel Prize Corpus: annotated with relation instances
  - Nobel Prize Corpus HPSG treebank (500 sentences) (resulted from the cooperation between Dan Flickinger and Peter Adolphs)

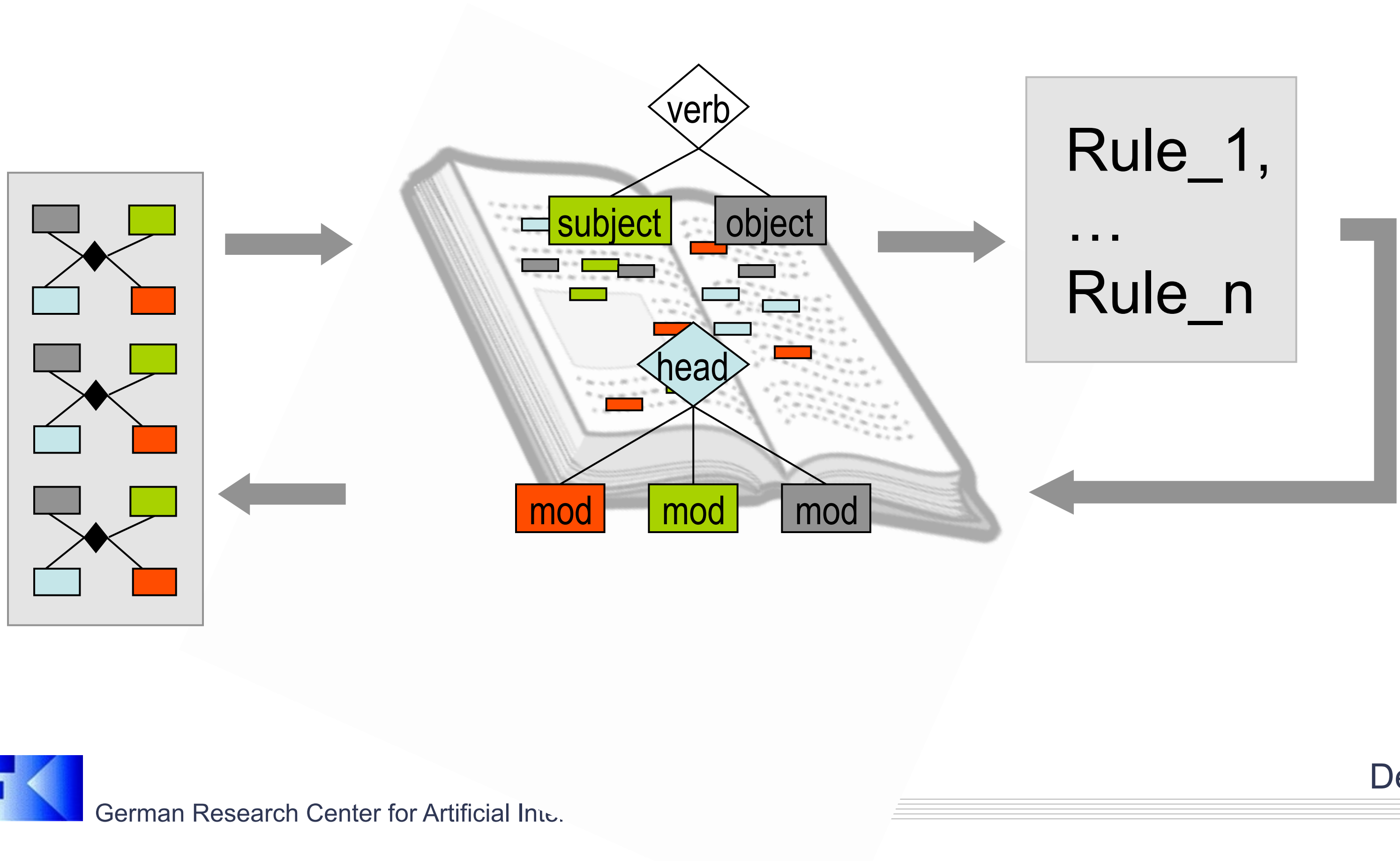
- ERG: 1004 release
- Redwood treebank (Oepen et al., 2002)
- *n*-best readings of parsing results
  - Parse selection model: a discriminative log-linear disambiguation model (Toutanova et al., 2005)

$$P(t|w) = \frac{\exp \sum_{i=1}^{n} \lambda_i f_i(t, w)}{\sum_{t' \in T(w)} \exp \sum_{i=1}^{n} \lambda_i f_i(t', w)} \qquad (1)$$

  - Best readings are decoded efficiently from a packed parse forest with dynamic programming (Zhang et al., 2007)

DARE (Xu et al., 2007; Xu 2007; Xu et al., 2008; Uszkoreit et al., 2009; Xu et al., 2010)

http://dare.dfki.de



Example in Nobel Prize Award Domain

- Seed example, an instance of target relation:

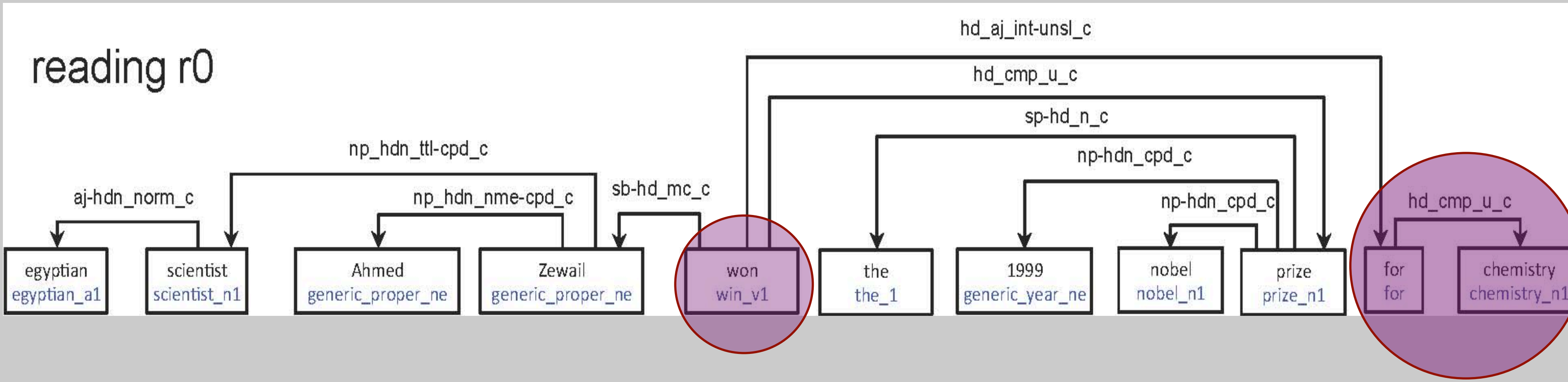
<Ahmed Zewail, Nobel, Chemistry, 1999>

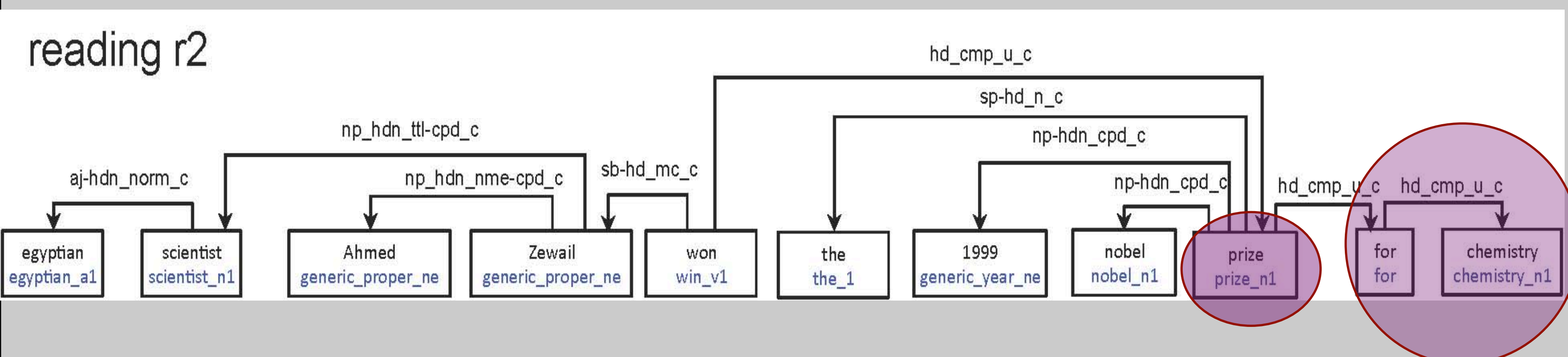
- DARE learns RE rules from parsing results of sentences which matched with the seed:

Egyptian scientist Ahmed Zewail won the 1999 Nobel Prize for Chemistry

HPSG Parses: PP Attachment

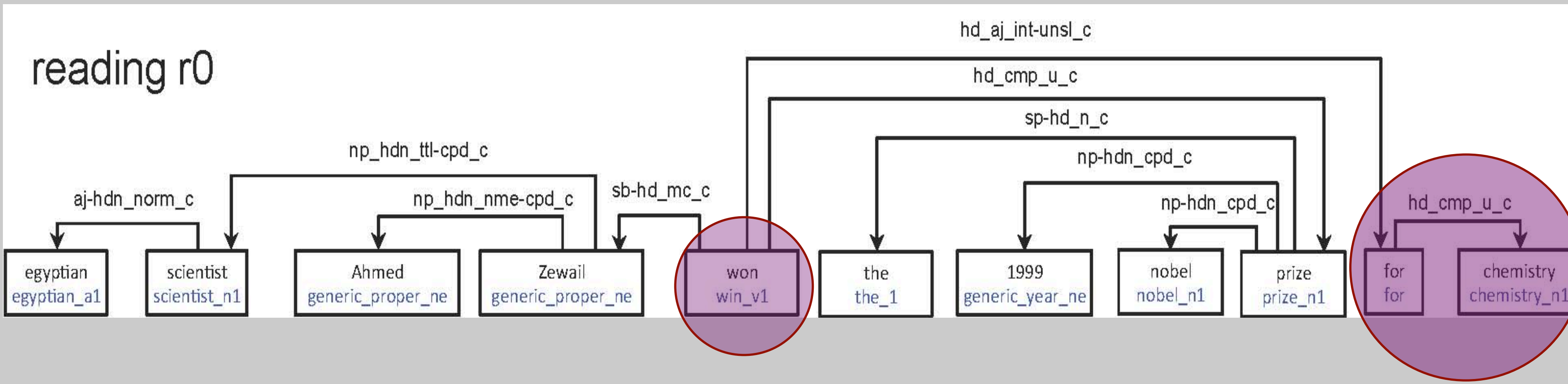
Egyptian scientist Ahmed Zewail won the 1999 Nobel Prize for chemistry

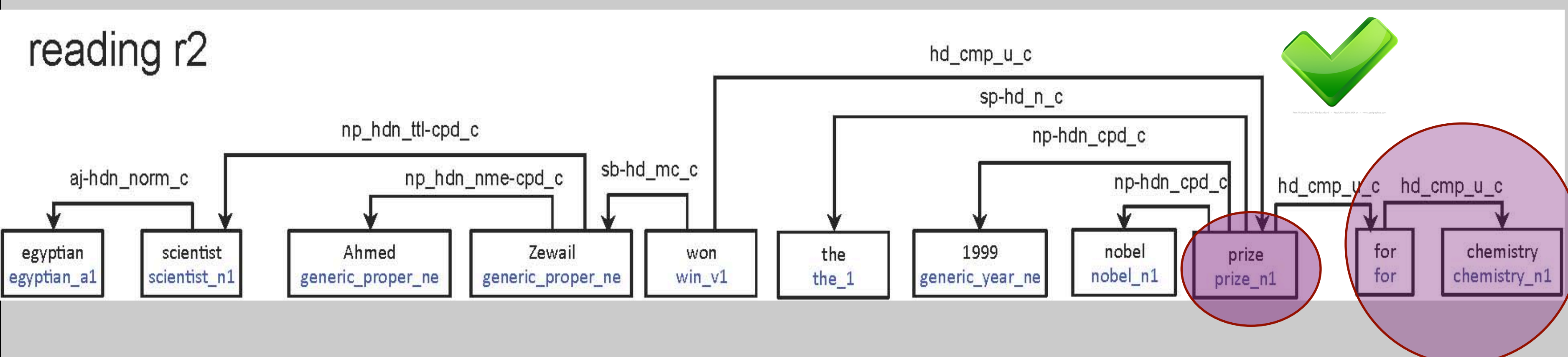




HPSG Parses: PP Attachment

Egyptian scientist Ahmed Zewail won the 1999 Nobel Prize for chemistry

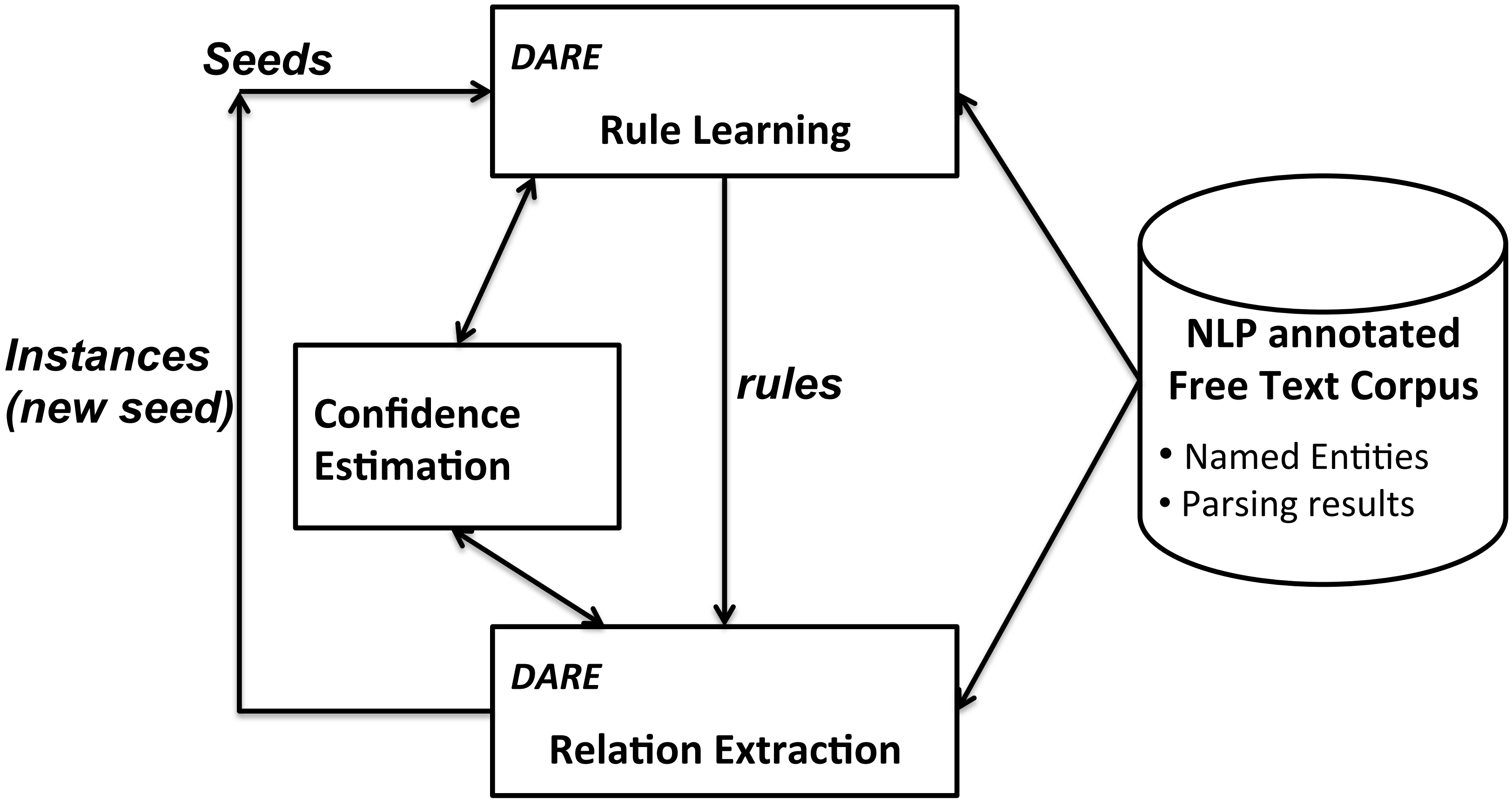




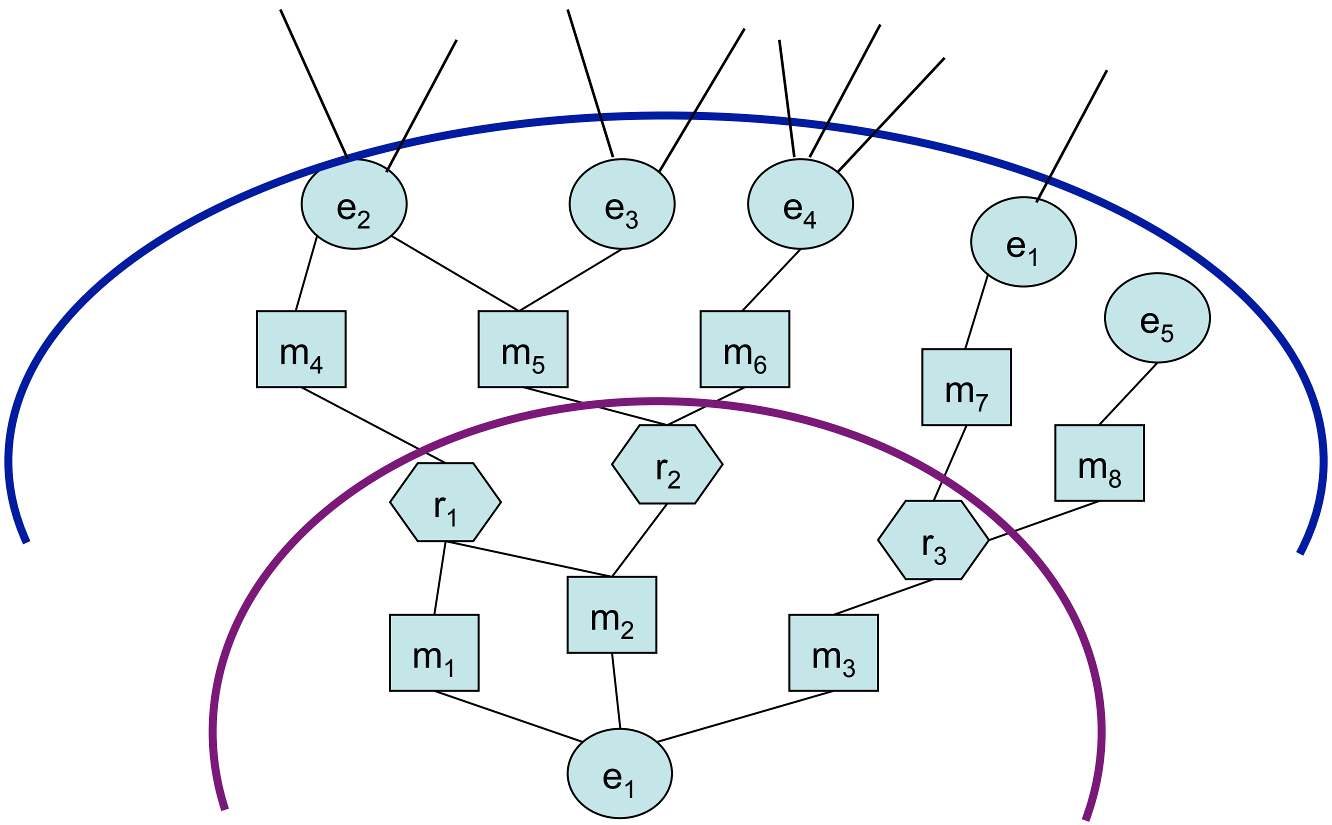
Rule_30 learned from *Reading R2*

rule_30
PATTERN
- *pattern*
- HEAD ("win_v1")
- SB-HD_MC_C
  - *sb-hd_mc_c*
  - HEAD $\boxed{0}$ <person>
- HD-CMP_U_C
  - *hd-cmp_u_c*
  - HEAD $\boxed{1}$ <prize>
  - HD-CMP_U_C
    - *hd-cmp_u_c_2*
    - HEAD ("for_prtcl")
    - HD-CMP_U_C
      - *hd-cmp_u_c_3*
      - HEAD $\boxed{2}$ <area>

OUTPUT
- *relation*
- area $\boxed{2}$
- winner $\boxed{0}$
- prize $\boxed{1}$


Seeds
DARE
Rule Learning
Instances
(new seed)
Confidence
Estimation
rules
NLP annotated
Free Text Corpus
• Named Entities
• Parsing results
DARE
Relation Extraction

Interaction of Rule Learning and Relation Extraction



- Duality principle (Brin, 1998;Yangarber, 2001 and Agichtein & Gravano, 2000)
  - Confidence values of the learned rules are dependent on the truth value of their extracted instances and on the seed instances from which they stem
  - Confidence values of an extracted instance makes use of the confidence value of its ancestor seed instances.

Given the scoring of instances

1) the confidence values of a rule is the average of score of all instances extracted by this rule or
2) the average score of seed instances from which this rule is learned

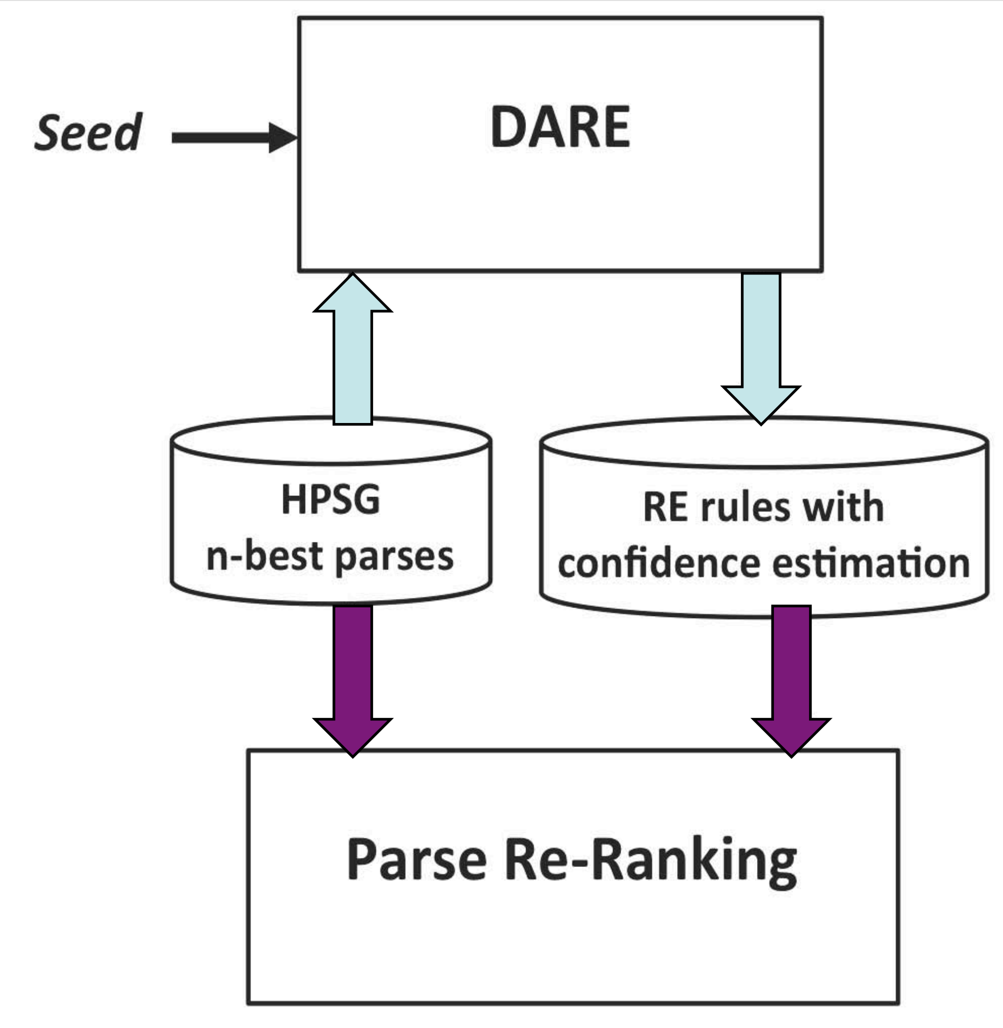
$$\mathbf{confidence}(rule) = \begin{cases} \dfrac{\sum_{i \in \mathbb{I}_{extracted}} \mathbf{score}(i)}{|\mathbb{I}_{extracted}|} & \textbf{if } \mathbb{I}_{extracted} \neq \phi \\[2ex] \dfrac{\sum_{j \in I_{rule}} \mathbf{score}(j)}{|I_{rule}|} \times \delta & \textbf{if } \mathbb{I}_{extracted} = \phi \end{cases}$$

$$\text{where} \quad \begin{aligned} &\mathbb{I}_{extracted} = \mathbf{getInstances}(rule), \\ &I_{rule} = \mathbf{getMotherInstancesOf}(rule), \\ &\delta = 0.5 \end{aligned}$$

In our reserach, we observe:

- A strong connection between RE task and the parser via the leared RE rules, because RE rules are derived from parses

- Confidence values of the RE rules imply the domain approriateness of the parse readings.



Parse Scoring with Confidence Values of RE Rules

$$S(t) = \begin{cases} \sum_{r \in R(t)} (\mathbf{confidence}(r) - \phi\mathbf{confidence}) & if\, R(t) \neq \phi, \\ 0 & if\, R(t) = \phi. \end{cases} \quad (6)$$

R(t): set of RE rules matching parse reading *t*, and
Φconfidence is the average confidence score among all rules.

The score of the reading will be increased if the matching rule has an above average confidence score.

Algorithm 1 compare_readings(r_i, r_j)

if compare($S(r_i), S(r_j)$) $\neq 0$ **then**
  return compare($S(r_i), S(r_j)$)
else # *Tie-breaking with MaxEnt scores*
  return compare($MaxEnt(r_i), MaxEnt(r_j)$)
end if

Data

- Nobel Prize corpus
  - 2864 documents from BBC, CNN and NYT: 143289 sentences
  - ERG covers 70% of the total corpus
- Gold-standard for evaluation
  - Nobel Prize corpus annotated with relation instances
  - 500 sentences of gold-standard HPSG treebank from Nobel Prize corpus

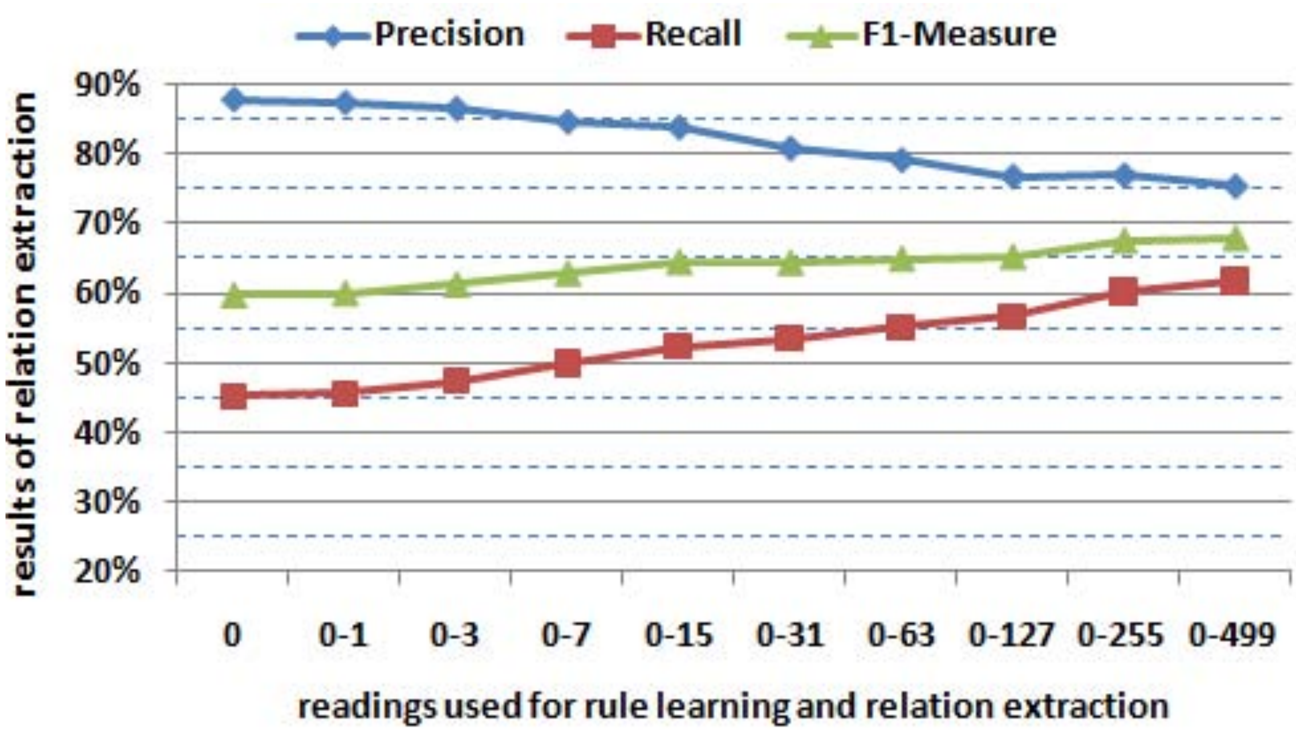
Experiments and Evaluation

- Training and test phases: RE performance
  - Baseline: without re-ranking
  - After re-ranking
- Qualitative analysis
  - Parsing performance after re-ranking
  - Rule quality after re-ranking

Training phase: learning re-ranking model

- We learn DARE rules from all 500 readings from all sentences in the training corpus.
- Given the rules and their confidence values, we re-rank the 500 readings of each sentence in the training corpus
- The re-ranking model is also applied to the test corpus

- Keep first *n=500* readings of all sentences and run DARE for rule learning and RE
- Observe whether correct relation instances can also be detected in the lower-ranked readings
  - Best reading: high precision, low recall, low F-measure
  - 500 readings: lower precision, higher recall, higher F-measure



- Training phase: evaluation
  - RE performance with the first reading before and after re-ranking

Reading 0	**Precision**	**Recall**	**F1-Measure**
Baseline (no re-ranking)	87.83%	45.18%	59.66%
After re-ranking	83.87%	56.19%	67.29%

Table 1: Training phase: Comparison of RE performance before and after re-ranking.

After Re-Ranking: Readings Matched with Learned Rules

Test phase: RE Performance before and after Re-ranking

Reading 0	**Precision**	**Recall**	**F1-Measure**
Baseline (no reranking)	82.93%	45.37%	58.56%
cwDB(after re-ranking)	80.33%	53.41%	64.16%

Table 2: Test phase: Comparison of RE performance before and after re-ranking.

Qualitative Analysis

- Experiments in both training and test phases confirm that our re-ranking improves recall and F-measure

- A further observation is that the ranked best readings are much more compatible with the learned DARE rules.

- Questions:
  - Whether re-ranking also improves parsing accuracy?
  - Whether a good reading for RE is also necessarily linguistically correct?

- We compare the syntactic structures against a high quality gold-standard treebank annotated by Dan Flickinger
  - Table 3 shows that the general parsing performance suffers from re-ranking both with respect to full trees and subtrees.

Model	$LB_{f_1}(full)$	$LB_{f_1}(subtree)$
MaxEnt	0.8613	0.8918
Reranked	0.7966	0.8132

Table 3: Labeled bracketing f-score

Error Analysis

- 113 test sentences, 68 show a different re-ranking
  - Improvement:
    - Labeled bracketing accuracy: 13
    - Better appositions: 3
    - Better selection of verb subcat frames: 2
    - Better PP attachments: 6
  - Degradation
    - Incorrect compounding in NPs: 24
    - Bad coordination: 7
    - Wrong lexical categories: 2

	"good" for RE
Before re-ranking	50%
After re-ranking	85%

Table 4: "Good" readings for RE among 68 re-ranked sentences

Explanation

- Linguistically „wrong“ analyses nevertheless lead to consistent extraction of rules and instances
- The increased consistency in the re-ranked parses does help improve the RE process.

For example: compound noun phrase: „Nobel Peace Prize laureate“

- Gold-standard bracketing: ((Nobel (Peace Prize)) laureate)
- Re-ranking reading: ((Nobel Peace) (Prize laureate))

The rule derived from the wrong reading can be applied to all equally incorrect readings of similar compound nouns:

“Nobel Chemistry/Physics/Economics Prize laureate”

Evaluation of Extraction Rules before and after Re-ranking

- The major contribution of re-ranking is not the improvement of general linguistic selection but the improvement of the selection of good readings for RE tasks
  - Good reading: rules learned from them extract correct instances
  - Bad reading: rules learned from them extract only incorrect instances
  - Useless reading: rules learned from them extract no instances

	Good Reading	Bad Reading	Useless Reading
before re-ranking	29.2%	1.3%	69.5%
after re-ranking	42.4%	0.8%	56.8%

Table 5: Test corpus: distribution of good readings before and after re-ranking

Conclusion

- The main contribution of our work is a method for adapting generic parsers to the tasks and domains of relation extraction by parse re-ranking.
- Our re-ranking is based on feedback from the application.
- We could show that for one generic parser/grammar, recall and f-measure could be considerably improved and hope that this effect can also be obtained for other generic parsers.
- Insights to share
  - Better parse ranking for the RE does not necessarily corresponds to a better parse ranking for other purposes or for generic parsing
  - The ease and consistency of rule extraction and rule application counts more than the linguistically correct analysis

Future Work

- The presented results may be viewed as a step forward toward making deep linguistic grammars useful for relation extraction
- Next steps will be dedicated to
  - Balancing off the deficits in coverage by
    - data-driven lexicon extension in the spirit of (Zhang et al., 2010) and
    - exploiting the chart for partial parses involving the relevant types of named entities
  - Application of our methods to other generic parsers
  - or whether the set of learned RE rules with their confidence values can be directly used as features in the statistical parse disambiguation models instead of in the post-processing step