

Hybrid Information Extraction

PD Dr. Günter Neumann
DFKI GmbH

Hybrid

- Is a system, if consists of different technologies
 - can be combined
 - each one depicts a solution by its own
 - the integration constitute an innovative plus for the whole system

Examples

Examples

hybrid engine

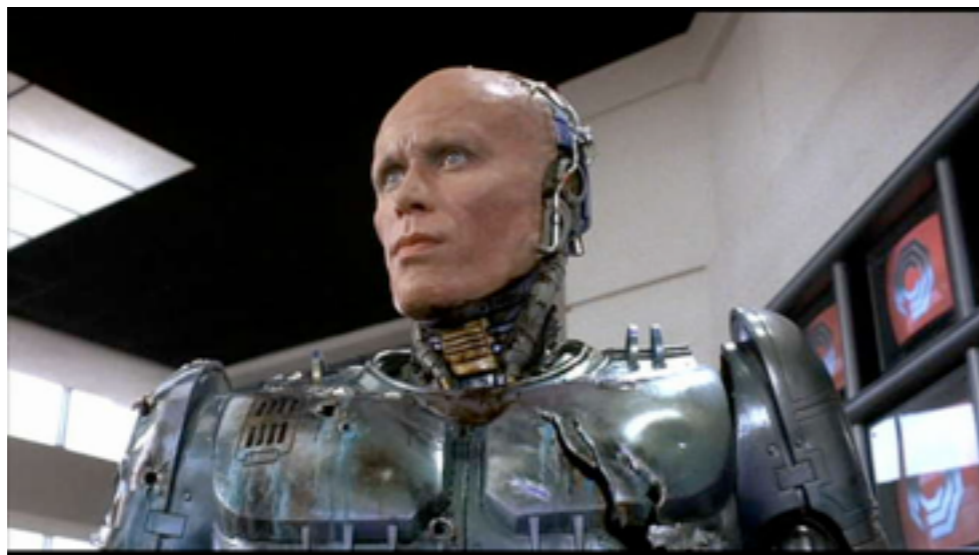


Examples

hybrid engine



HumanMachine

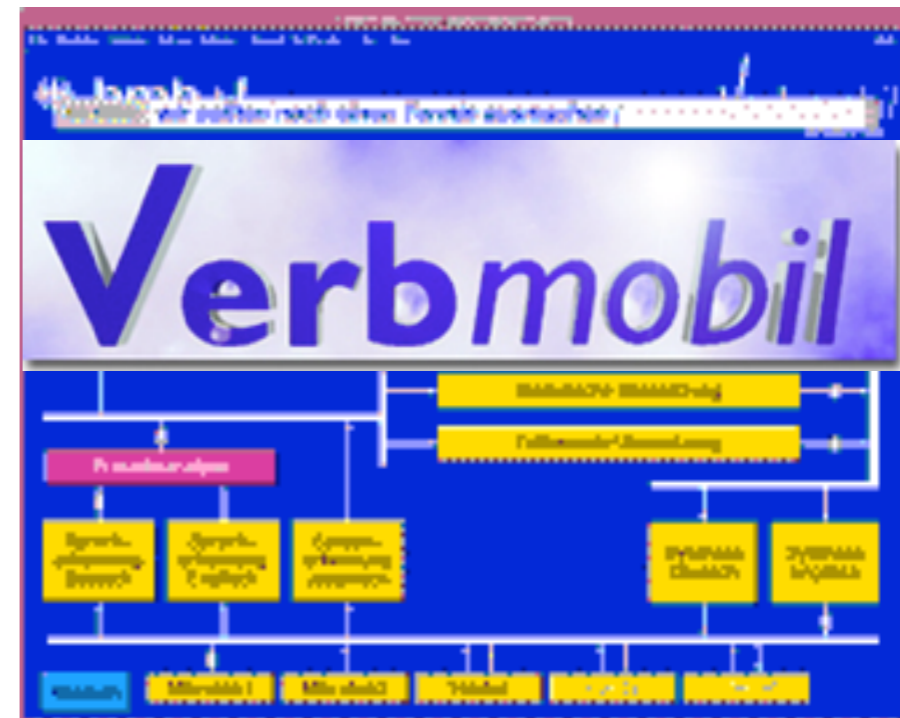


Examples

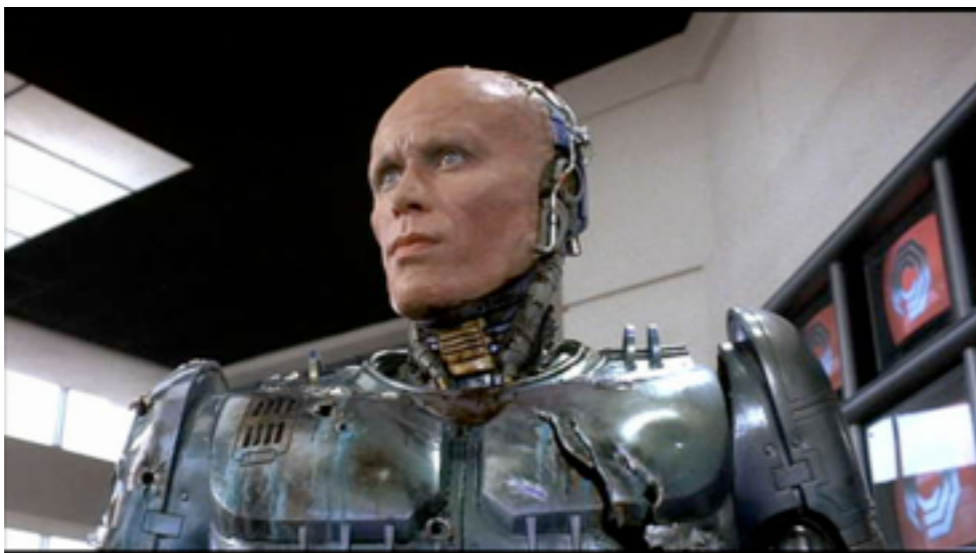
hybrid engine



Hybrid Language Processing



HumanMachine



Information Extraction

- The aim of information extraction (IE) is the identification and structuring of domain specific information from free text by skipping irrelevant information at the same time.
- What counts as relevant is given to the system in form of pre-defined domain specific annotations, lexicon entries or rules.

Example: news about turnover

Example: news about turnover

turnover(Company, Year, Manner, Amount, Tendendcy, Differnce)

Unternehmen	Jahr	Größe	Betrag	Tendenz	Differenz
Compaq	1998	Umsatz	31 Mrd. USD	+	27%

Example: news about turnover

turnover(Company, Year, Manner, Amount, Tendendcy, Differnce)

Unternehmen	Jahr	Größe	Betrag	Tendenz	Differenz
Compaq	1998	Umsatz	31 Mrd. USD	+	27%

Eine Mixtur aus wachsendem Dienstleistungsgeschäft, Kostensenkungen und erfolgreichen Akquisitionen brachte Wettbewerber IBM im zweiten Quartal deutlich verbesserte Ergebnisse. Zwischen April und Juni stiegen der Umsatz um 10% auf 21,6 Mrd.\$ und der Reingewinn auf 1,7 Mrd.\$. Sonderlasten in Höhe von 1,4 Mrd.\$ hatten den Vorjahresgewinn auf 56 Mill.\$ gedrückt.

Example: news about turnover

turnover(Company, Year, Manner, Amount, Tendendcy, Differnce)

Unternehmen	Jahr	Größe	Betrag	Tendenz	Differenz
Compaq	1998	Umsatz	31 Mrd. USD	+	27%

Eine Mixtur aus wachsendem Dienstleistungsgeschäft, Kostensenkungen und erfolgreichen Akquisitionen brachte Wettbewerber IBM im zweiten Quartal deutlich verbesserte Ergebnisse. Zwischen April und Juni stiegen der Umsatz um 10% auf 21,6 Mrd.\$ und der Reingewinn auf 1,7 Mrd.\$ Sonderlasten in Höhe von 1,4 Mrd.\$ hatten den Vorjahresgewinn auf 56 Mill.\$ gedrückt.

Unternehmen	Jahr	Größe	Betrag	Tendenz	Differenz
IBM	2003	Umsatz	21,6 Mrd. \$	+	10 %

Example: news about turnover

turnover(Company, Year, Manner, Amount, Tendendcy, Differnce)

Unternehmen	Jahr	Größe	Betrag	Tendenz	Differenz
Compaq	1998	Umsatz	31 Mrd. USD	+	27%

Eine Mixtur aus wachsendem Dienstleistungsgeschäft, Kostensenkungen und erfolgreichen Akquisitionen brachte Wettbewerber IBM im zweiten Quartal deutlich verbesserte Ergebnisse. Zwischen April und Juni stiegen der Umsatz um 10% auf 21,6 Mrd.\$ und der Reingewinn auf 1,7 Mrd.\$. Sonderlasten in Höhe von 1,4 Mrd.\$ hatten den Vorjahresgewinn auf 56 Mill.\$ gedrückt.

Unternehmen	Jahr	Größe	Betrag	Tendenz	Differenz
IBM	2003	Umsatz	21,6 Mrd. \$	+	10 %

IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis

IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis

The founder of Microsoft, Bill Gates, lives in Seattle, Washington, which is also the place of the company's headquarter.

IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis

Bill Gates
is a Person



The founder of Microsoft, Bill Gates, lives in Seattle, Washington, which is also the place of the company's headquarter.

IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis

Bill Gates
is a Person



The founder of Microsoft, Bill Gates, lives in Seattle, Washington, which is also the place of the company's headquarter.



IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis

Bill Gates
is a Person



The founder of Microsoft, Bill Gates, lives in Seattle, Washington, which is also the place of the company's headquarter.

IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis

Seattle is a Location



Bill Gates
is a Person



founder_of



Microsoft

Microsoft
is a Company

The founder of Microsoft, Bill Gates, lives in Seattle, Washington, which is also the place of the company's headquarter.

IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis

Seattle is a Location



lives_in

Bill Gates
is a Person



founder_of



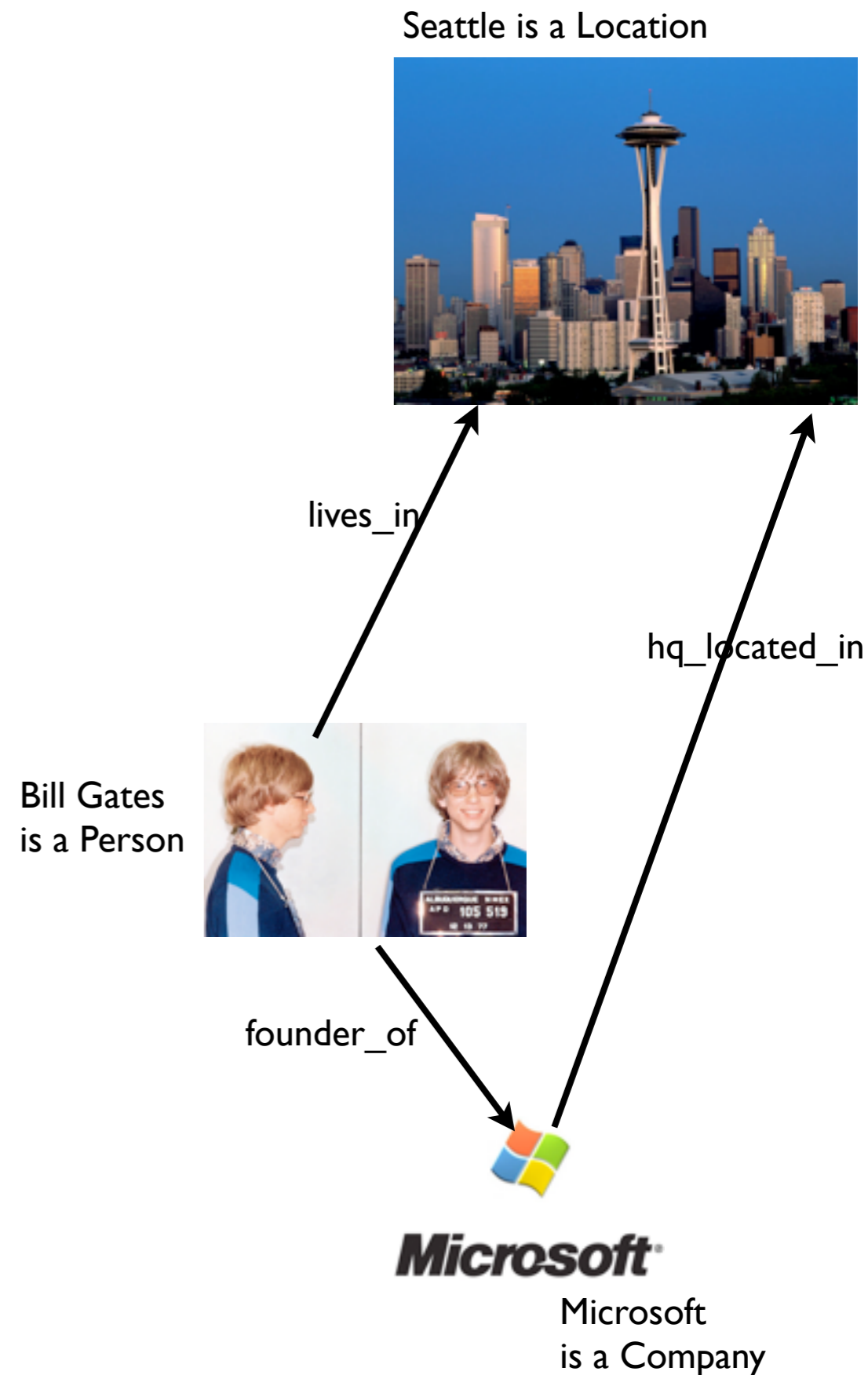
Microsoft

Microsoft
is a Company

The founder of Microsoft, Bill Gates, lives in Seattle, Washington, which is also the place of the company's headquarter.

IE - History

- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
 - Named-Entity Extraction (NER)
 - Relation Entity Extraction (REE)
 - Event Entity Extraction (EEE)
 - Coreferential analysis



The founder of Microsoft, Bill Gates, lives in Seattle, Washington, which is also the place of the company's headquarter.

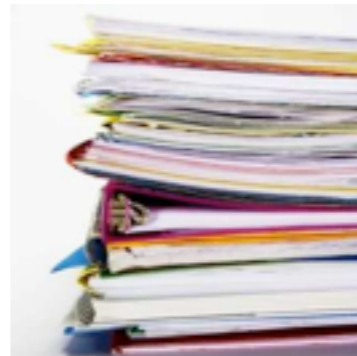
IE - the Present

- There exists knowledge-based IE (KIE) and statistical IE (SIE)
- SIE is the State-of-the-Art in research, WIE in industry
- There exists a number of different strategies for the various IE-subtasks
 - from simple gazetteers to complex ontologies
 - from supervised, to minimal supervised to unsupervised Machine Learning algorithms
- Recently, the research focus is on NER, REE, Web-based IE, scalability, domain adaptivity, ...
- Open question: Which method is actually better suited for which text source, domain and application?

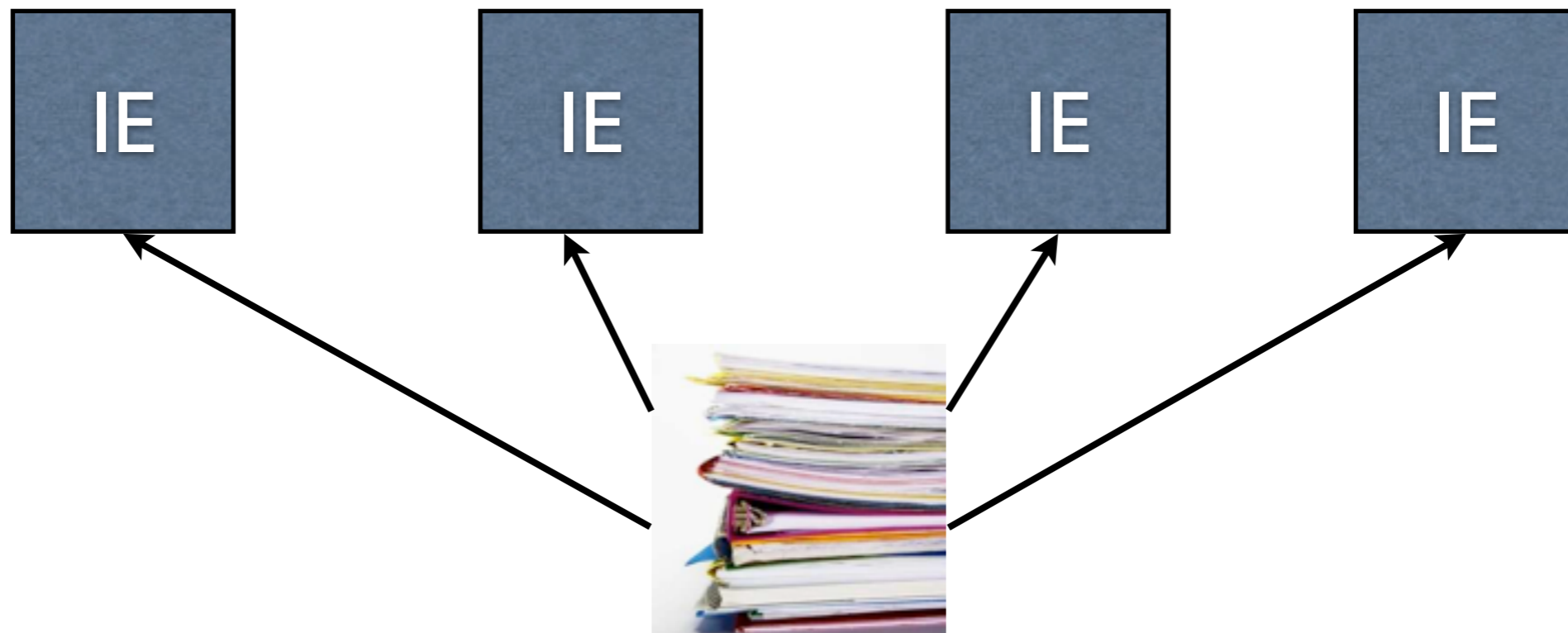
Hybrid IE

- Methods and strategies for the combination of different IE-components and the analysis of their plausibility.
- What are possible combinations ?

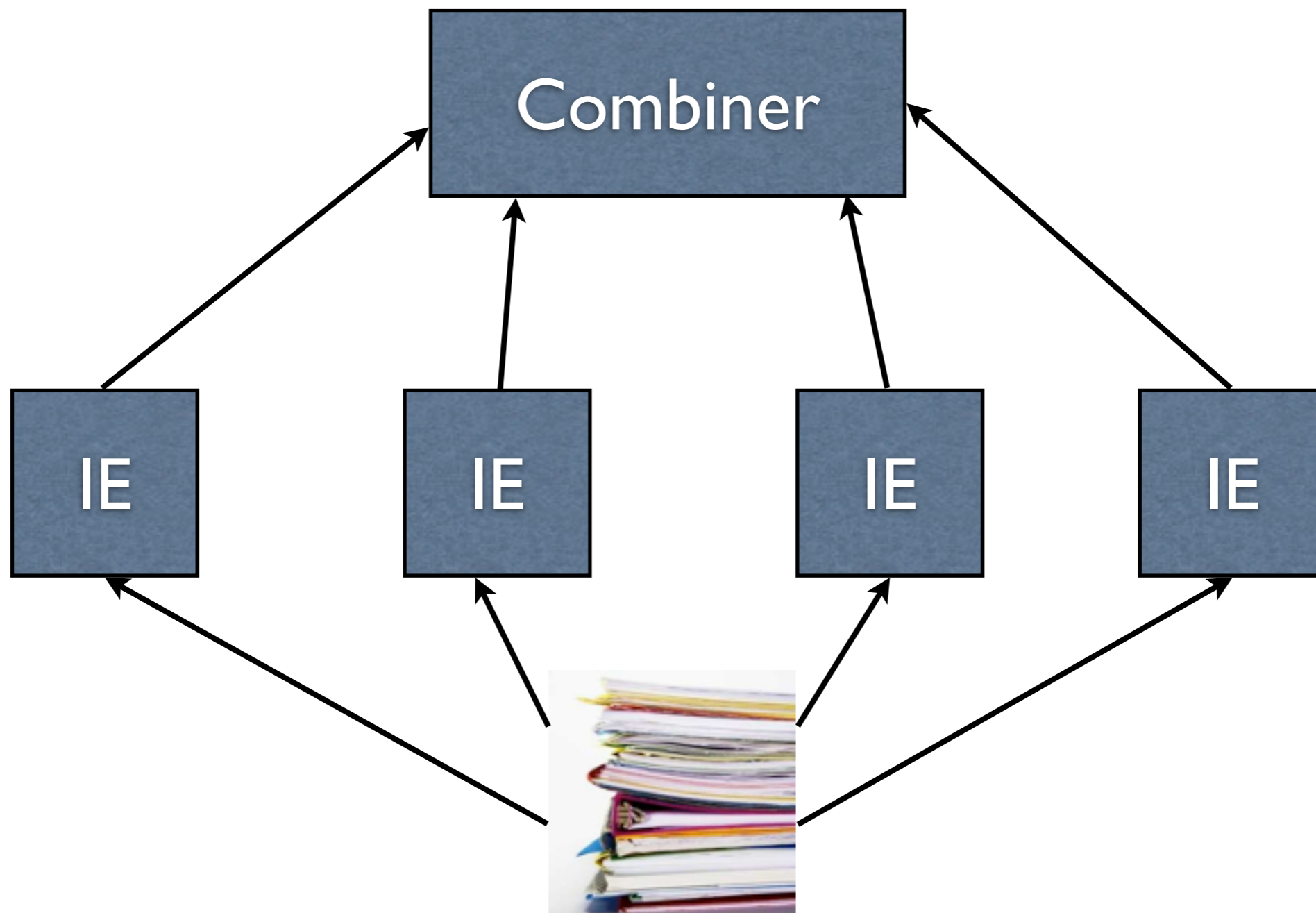
Multi-Strategy



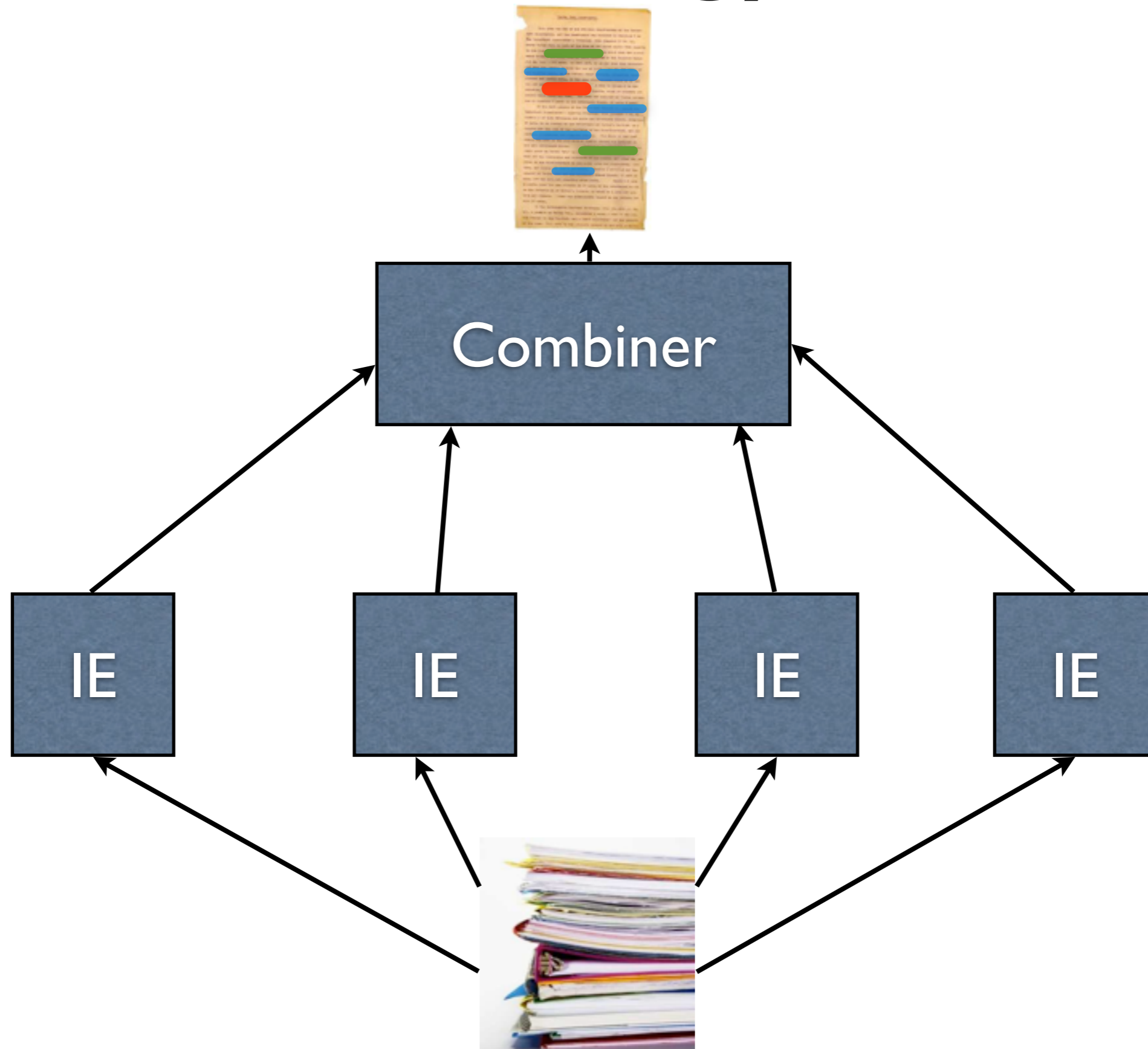
Multi-Strategy



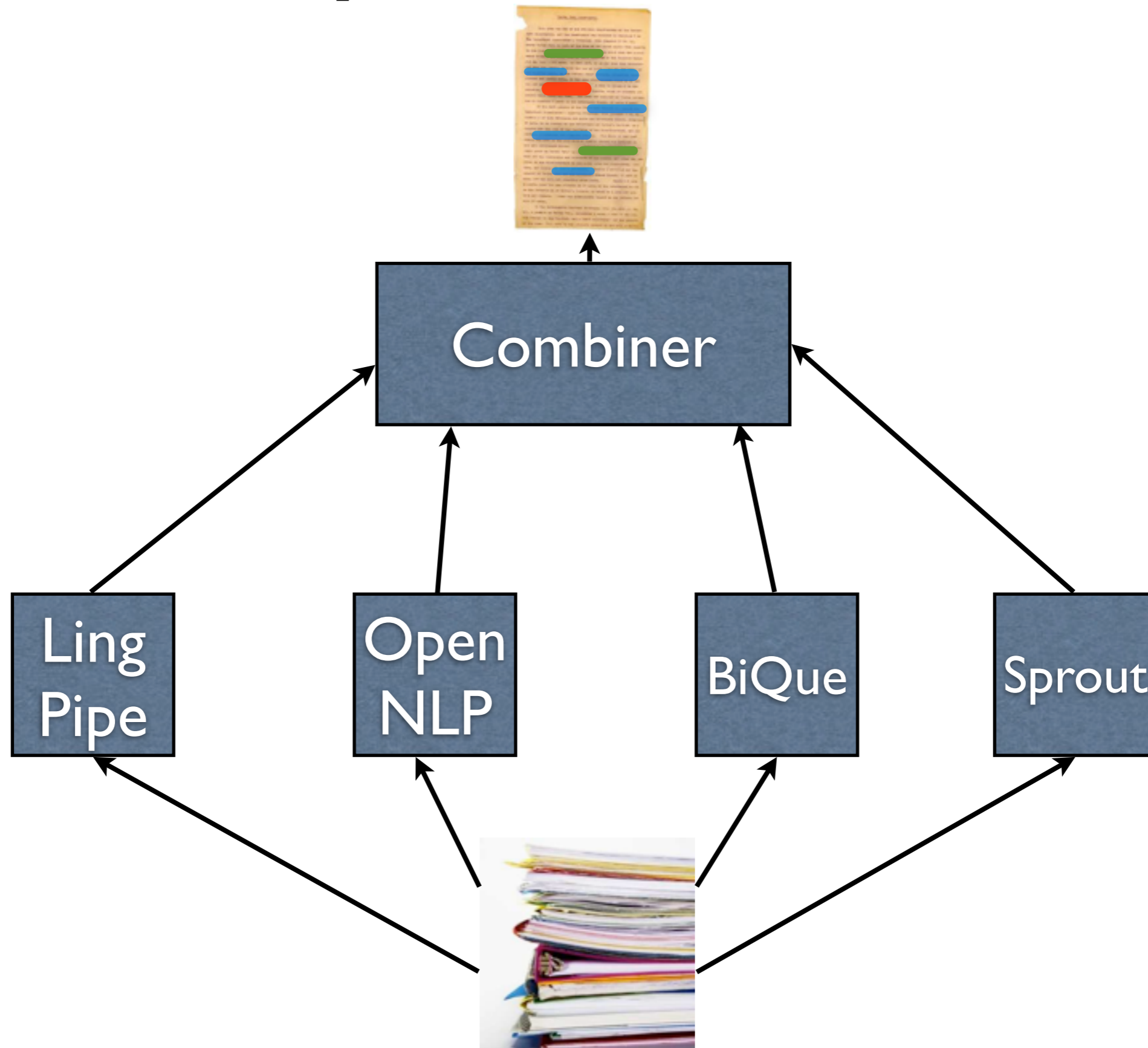
Multi-Strategy



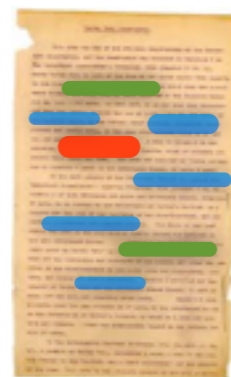
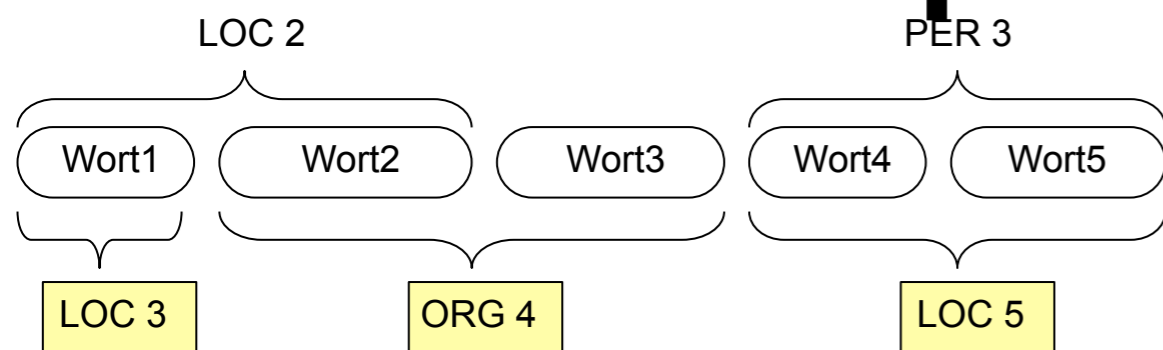
Multi-Strategy



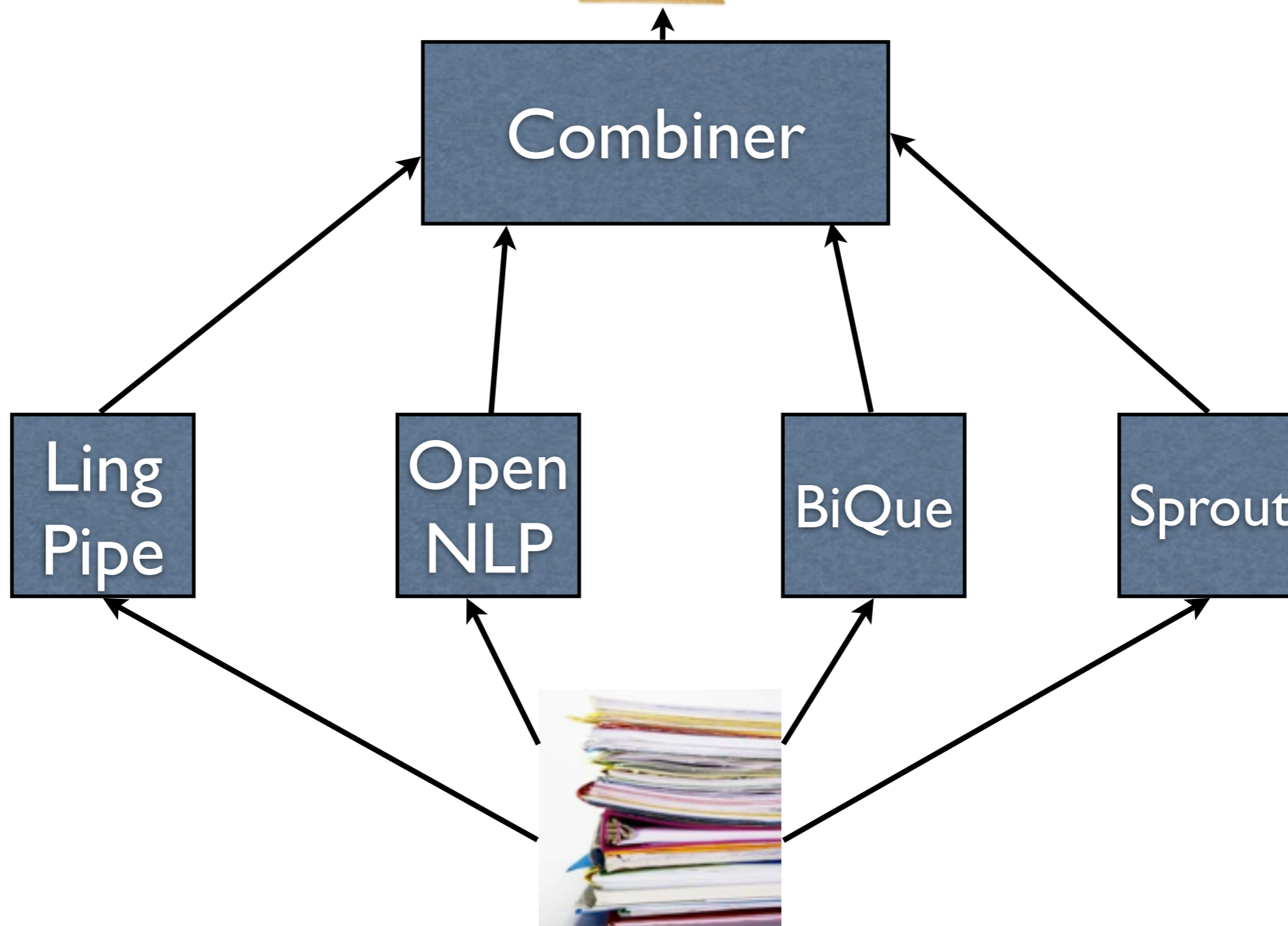
Example: NER



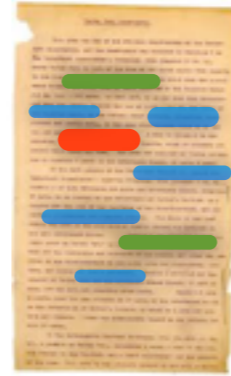
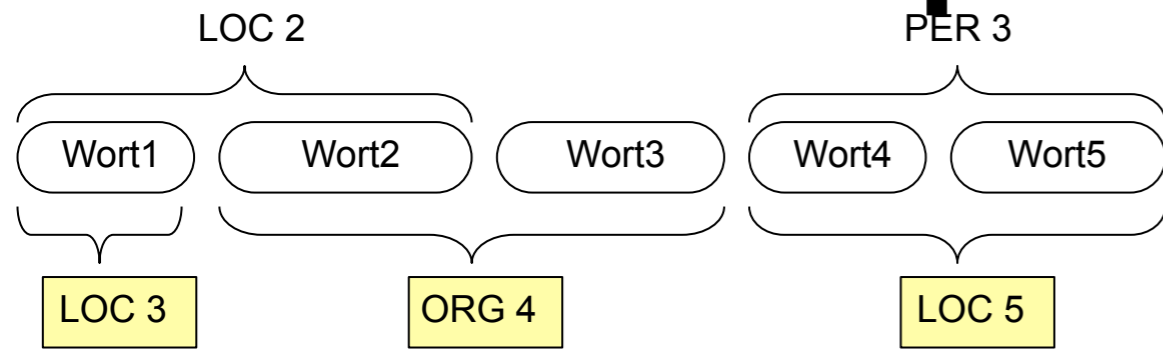
Example: NER



Problem:
- Ambiguities
- Bracketing



Example: NER

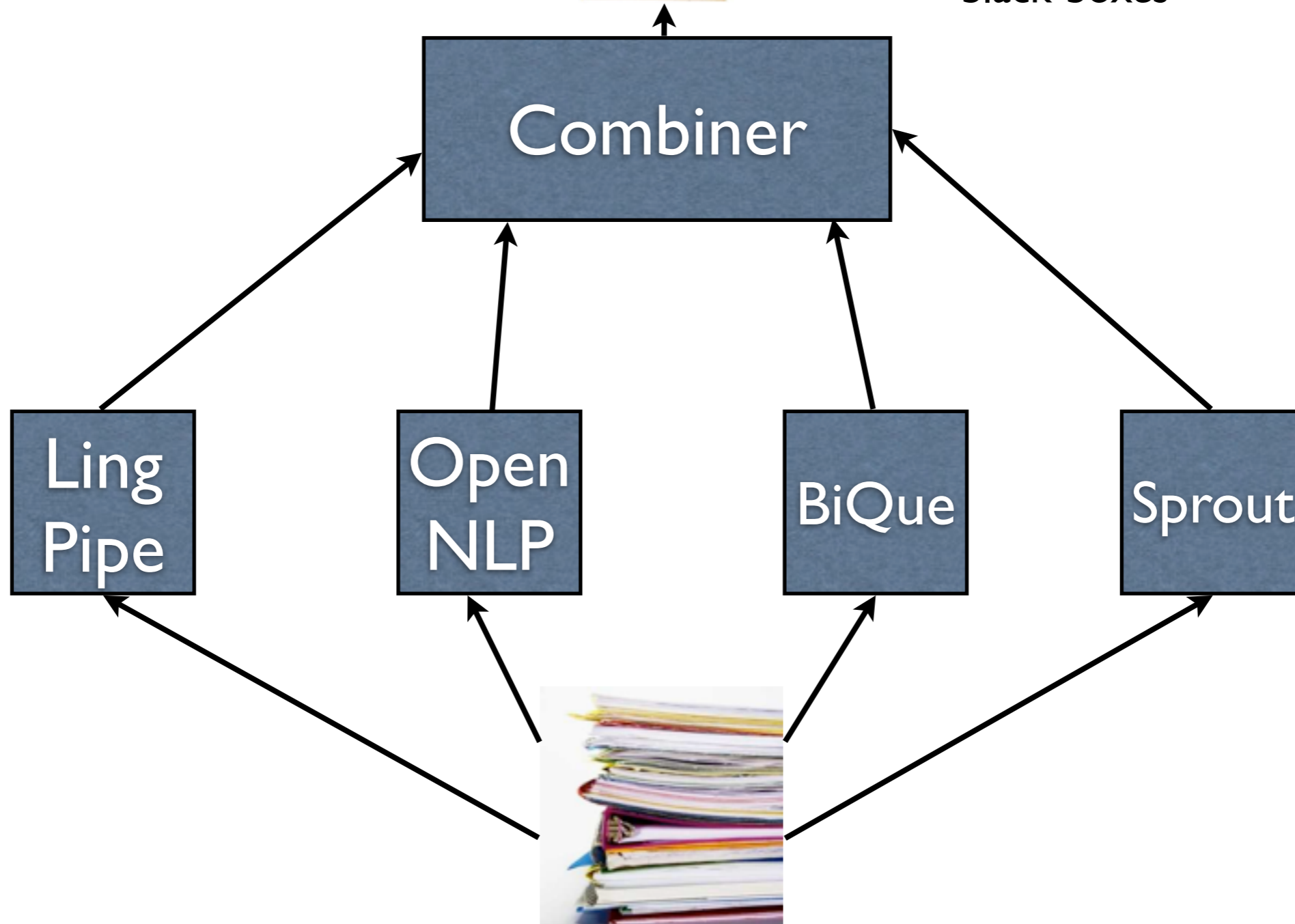


Solutions:

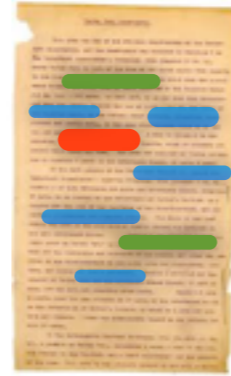
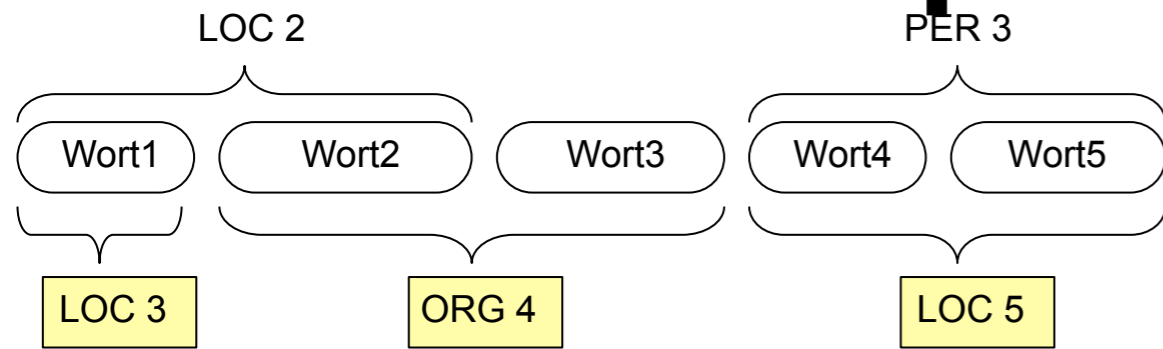
- meta-learning
- consider IE_i as independent black-boxes

Problem:

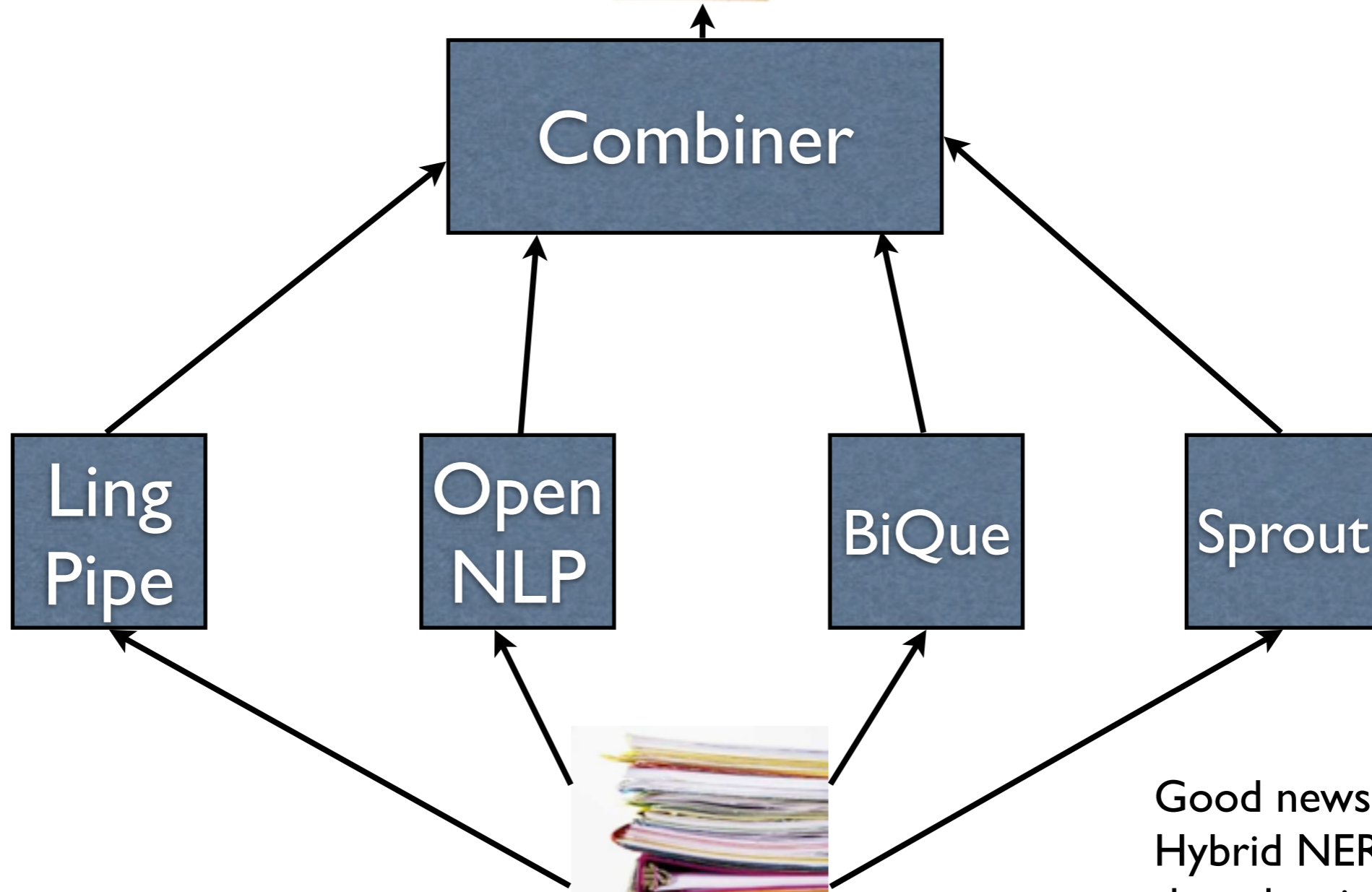
- Ambiguities
- Bracketing



Example: NER



Problem:
- Ambiguities
- Bracketing

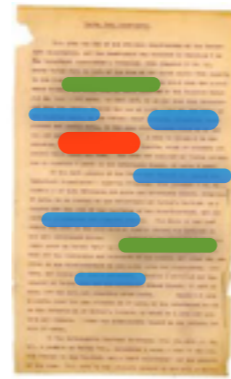
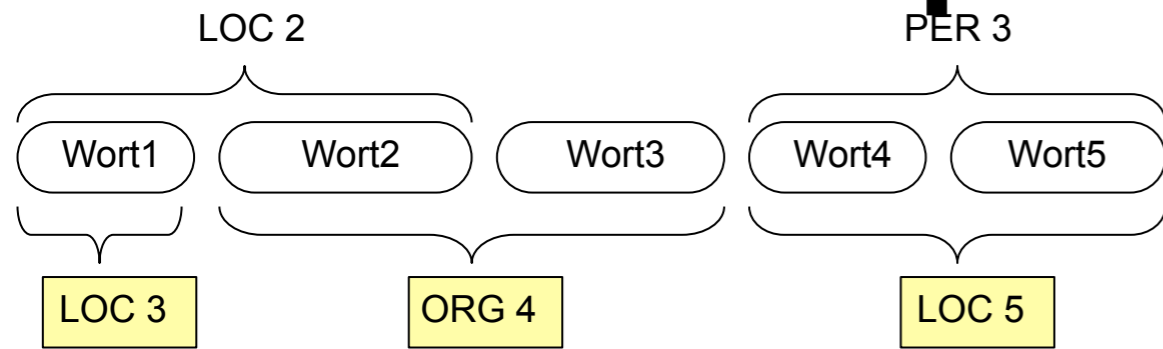


Good news:*

Hybrid NER are better than the single NER wrt. recall and precision.

Combining Information Extraction Systems Using Voting and Stacked Generalization
by: G Sigletos et al., J. Mach. Learn. Res., Vol. 6 (2005), pp. 1751-1782.

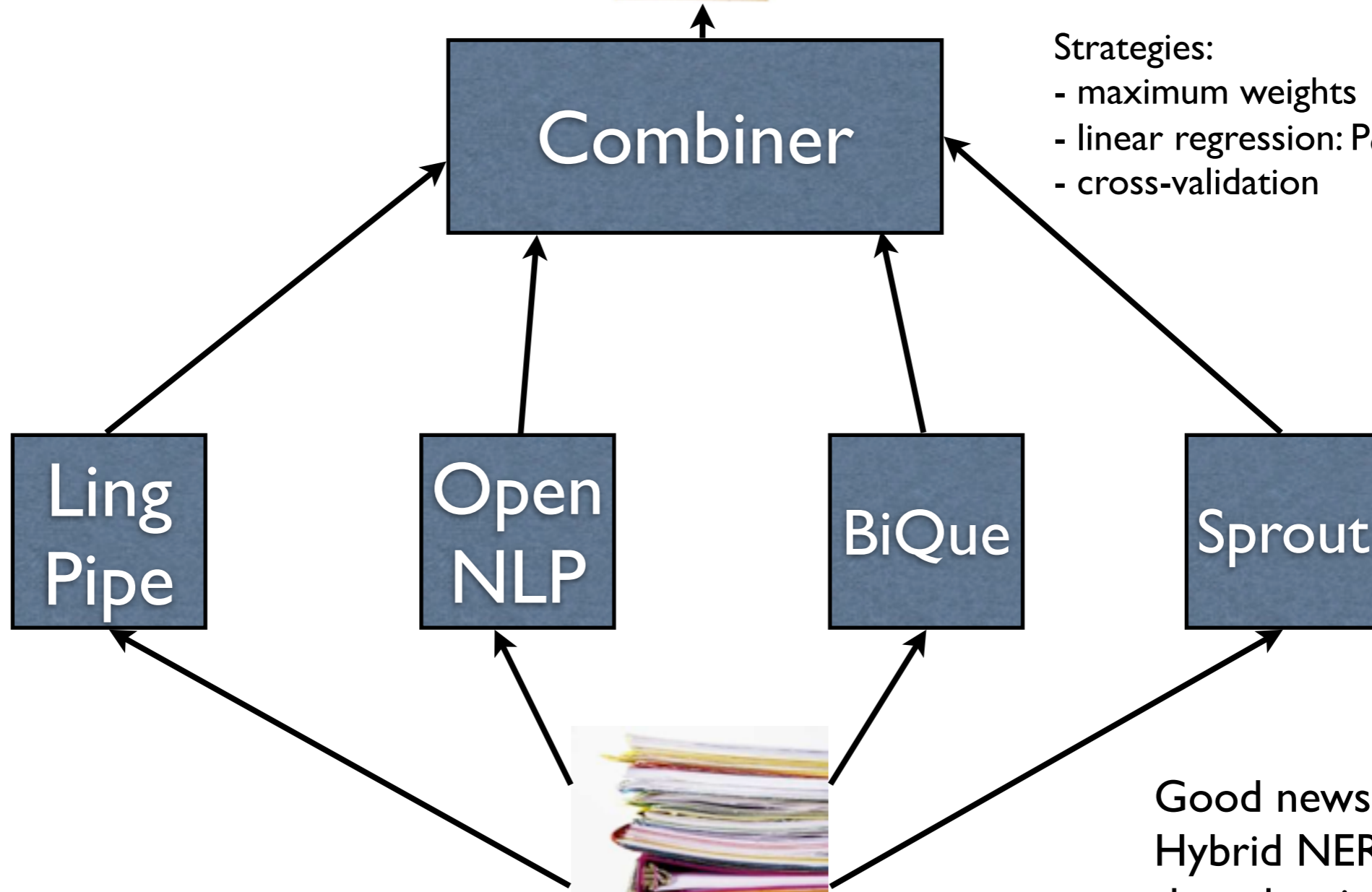
Example: NER



Meta learning
- majority voting
- stacking

Problem:
- Ambiguities
- Bracketing

Strategies:
- maximum weights
- linear regression: $P_C = 1 - \prod_i (1 - P_i)$
- cross-validation



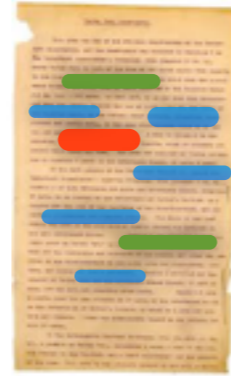
Good news:*

Hybrid NER are better than the single NER wrt. recall and precision.

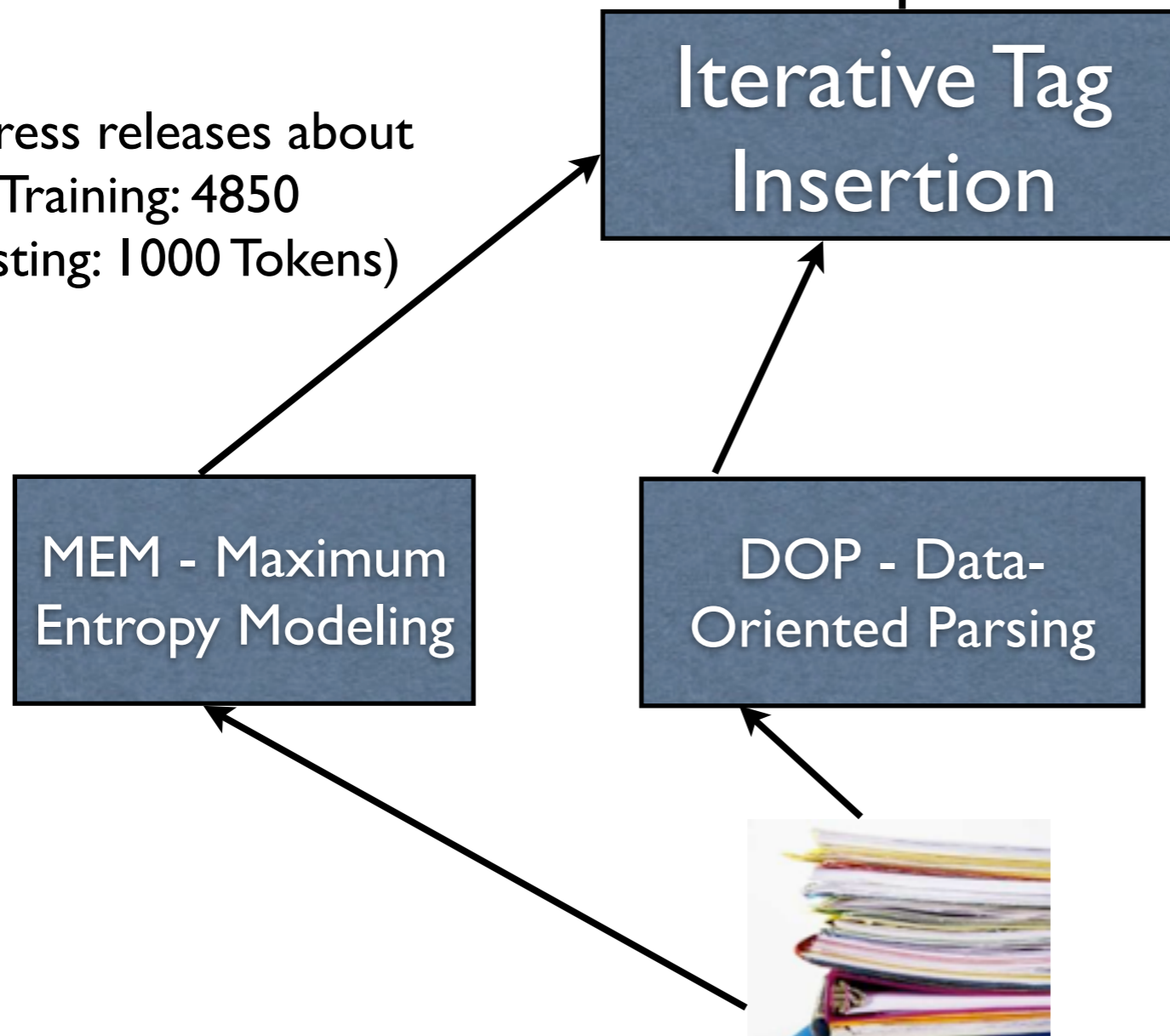
Combining Information Extraction Systems Using Voting and Stacked Generalization
by: G Sigletos et al., J. Mach. Learn. Res., Vol. 6 (2005), pp. 1751-1782.

Example: Template Filling

Der Gewinn <Org>der Schweppes GmbH & Co.</Org> KG
betrug <TIMEX>im ersten Quartal 1997</TIMEX> weit ueber 20 Mio. DM.



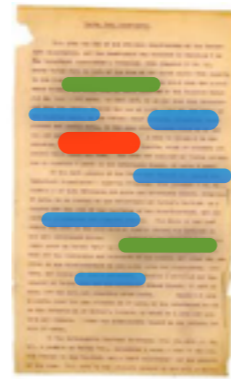
Corpus:
German press releases about turnover (Training: 4850 Tokens, Testing: 1000 Tokens)



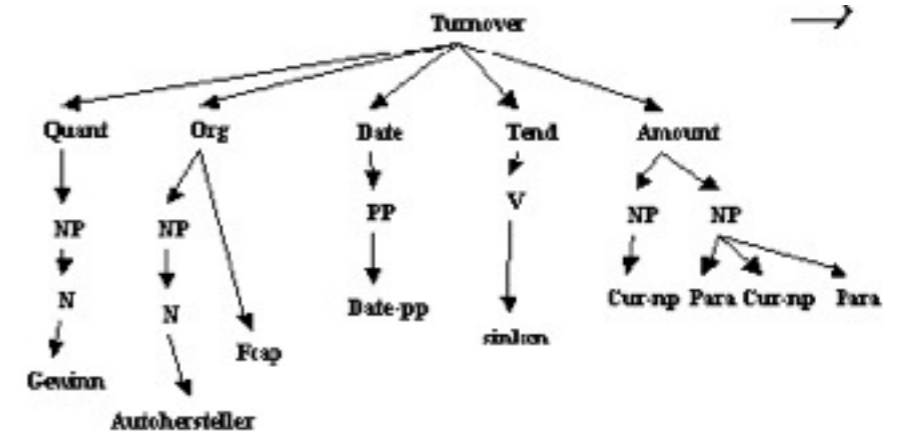
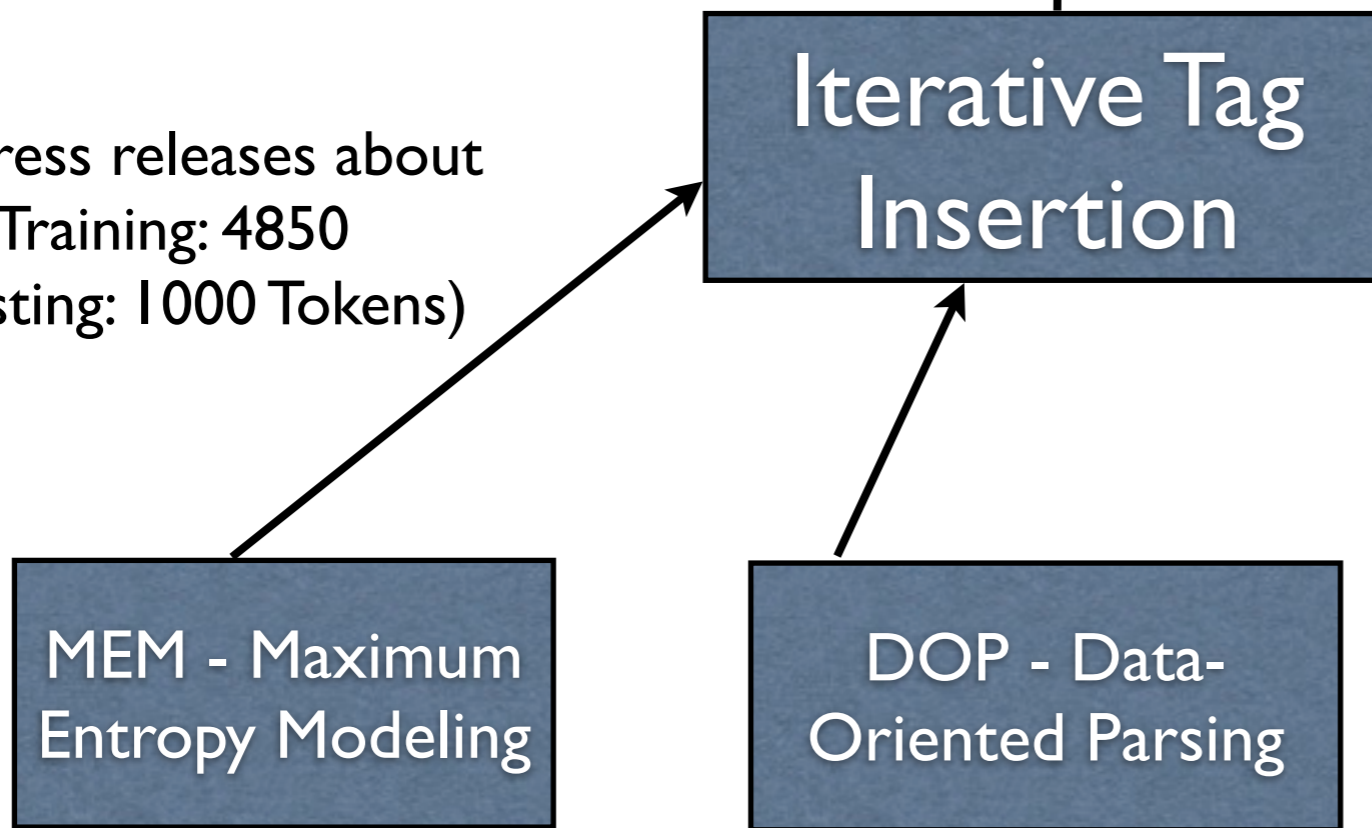
Neumann, G. (2006) A Hybrid Machine Learning Approach for Information Extraction from Free Texts. In Spiliopoulou et al. (Eds). From Data and Information Analysis to Knowledge Engineering, Springer series: Studies in Classification, Data Analysis, and Knowledge Organization, pages 390-397, Springer-Verlag Berlin, Heidelberg, New-York.

Example: Template Filling

Der Gewinn <Org>der Schweppes GmbH & Co.</Org> KG
betrug <TIMEX>im ersten Quartal 1997</TIMEX> weit ueber 20 Mio. DM.



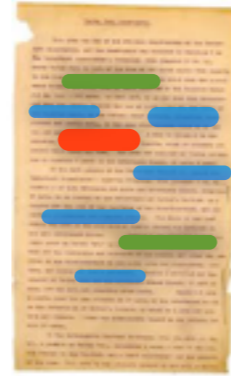
Corpus:
German press releases about turnover (Training: 4850 Tokens, Testing: 1000 Tokens)



Neumann, G. (2006) A Hybrid Machine Learning Approach for Information Extraction from Free Texts. In Spiliopoulou et al. (Eds). From Data and Information Analysis to Knowledge Engineering, Springer series: Studies in Classification, Data Analysis, and Knowledge Organization, pages 390-397, Springer-Verlag Berlin, Heidelberg, New-York.

Example: Template Filling

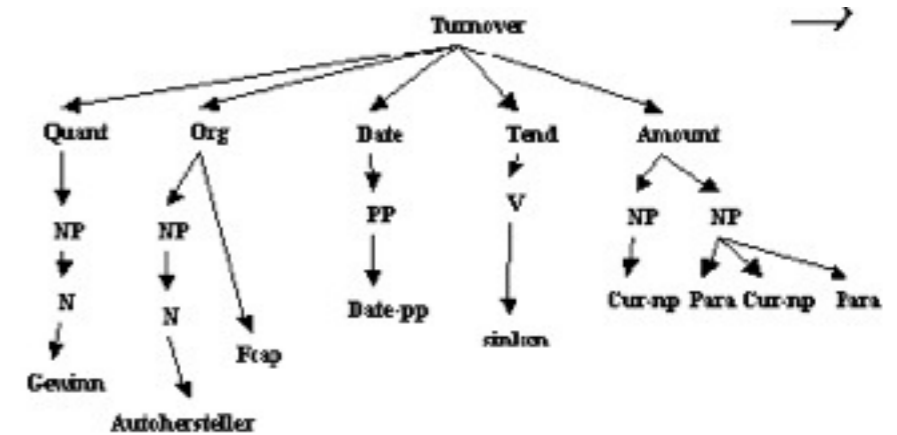
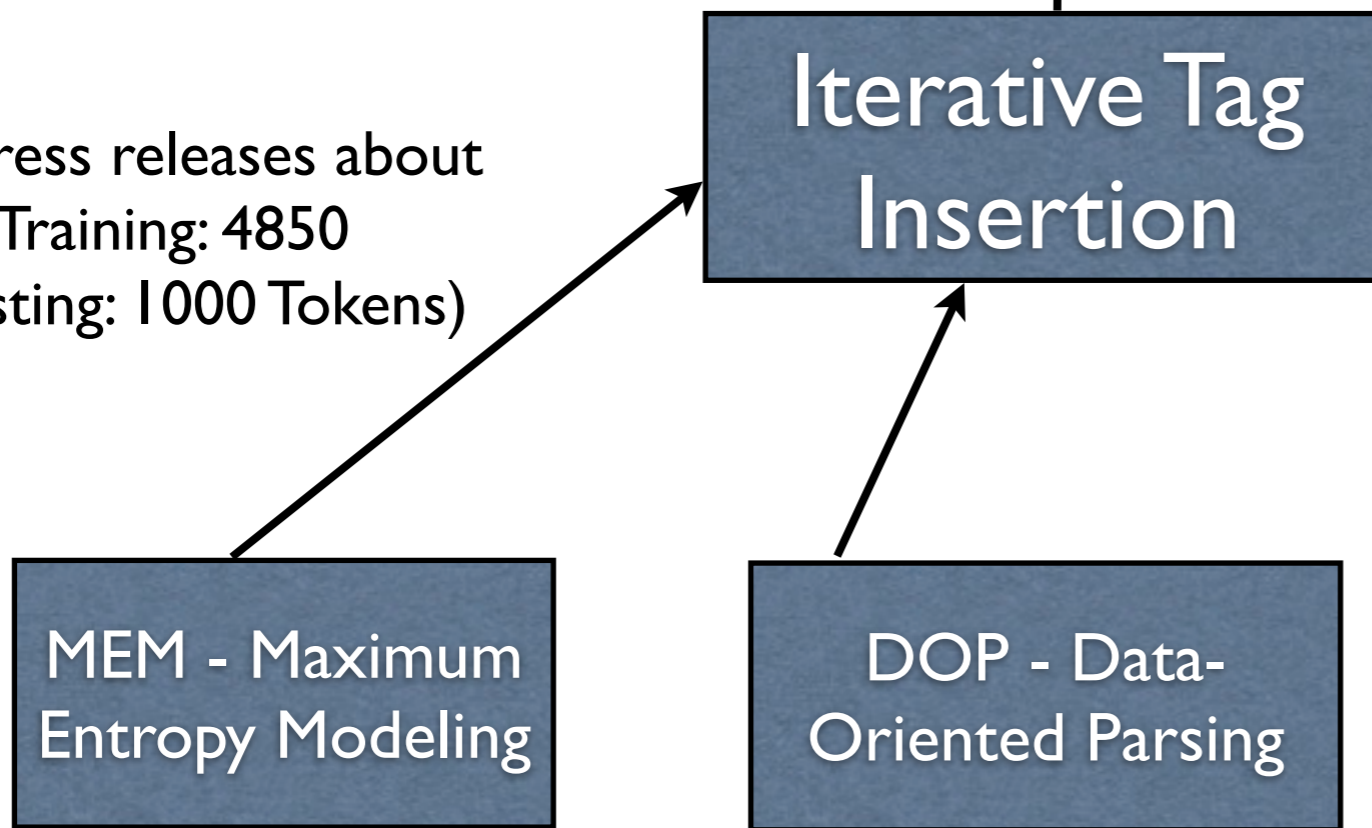
Der Gewinn <Org>der Schweppes GmbH & Co.</Org> KG
betrug <TIMEX>im ersten Quartal 1997</TIMEX> weit ueber 20 Mio. DM.



Result:

- only MEM: 79.3 %
- only DOP: 51.9 %
- both: 85.2 %

Corpus:
German press releases about turnover (Training: 4850 Tokens, Testing: 1000 Tokens)

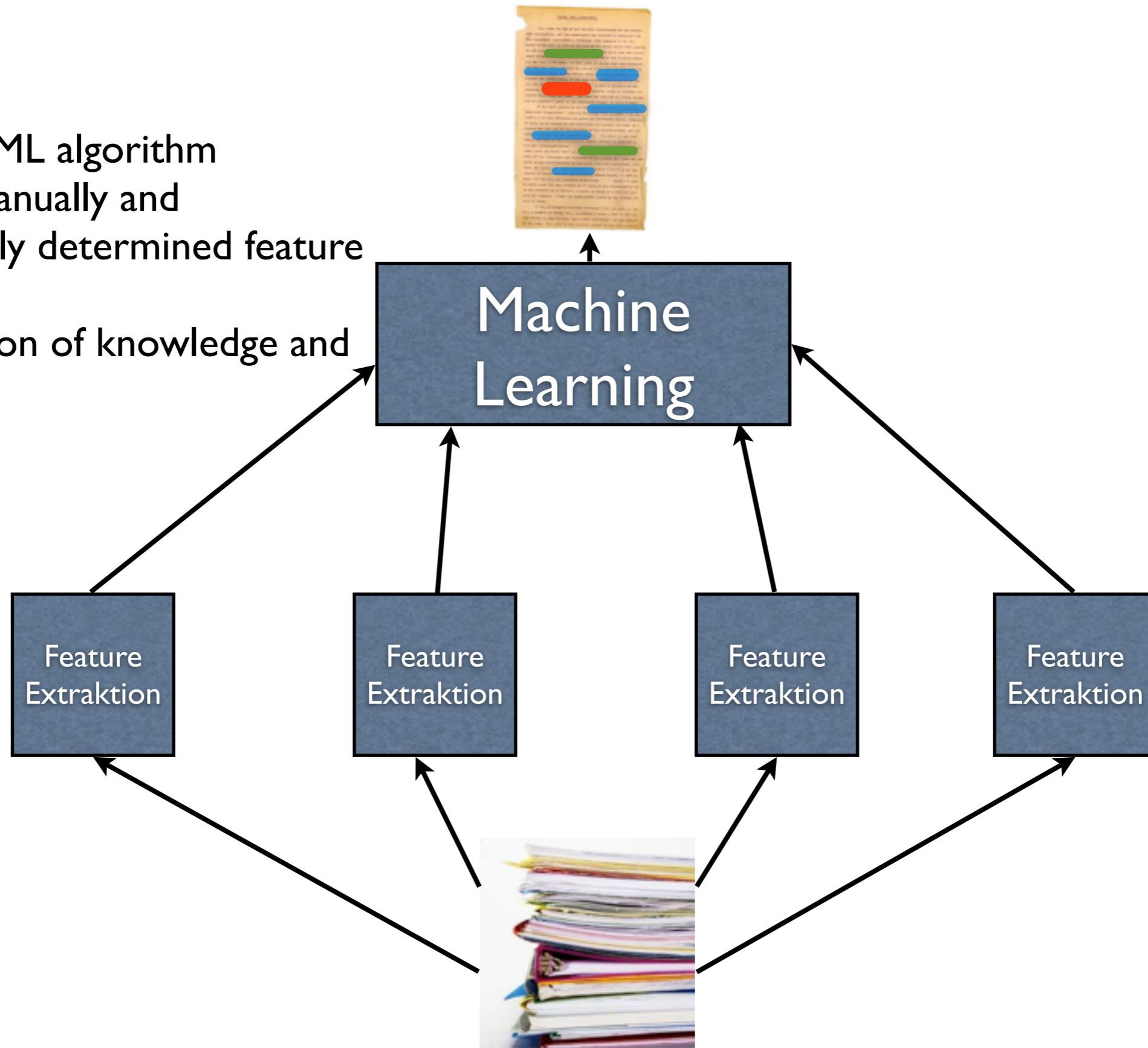


Neumann, G. (2006) A Hybrid Machine Learning Approach for Information Extraction from Free Texts. In Spiliopoulou et al. (Eds). From Data and Information Analysis to Knowledge Engineering, Springer series: Studies in Classification, Data Analysis, and Knowledge Organization, pages 390-397, Springer-Verlag Berlin, Heidelberg, New-York.

Feature based Strategies

Idea:

- choose a ML algorithm
- choose manually and automatically determined feature templates
- combination of knowledge and statistics

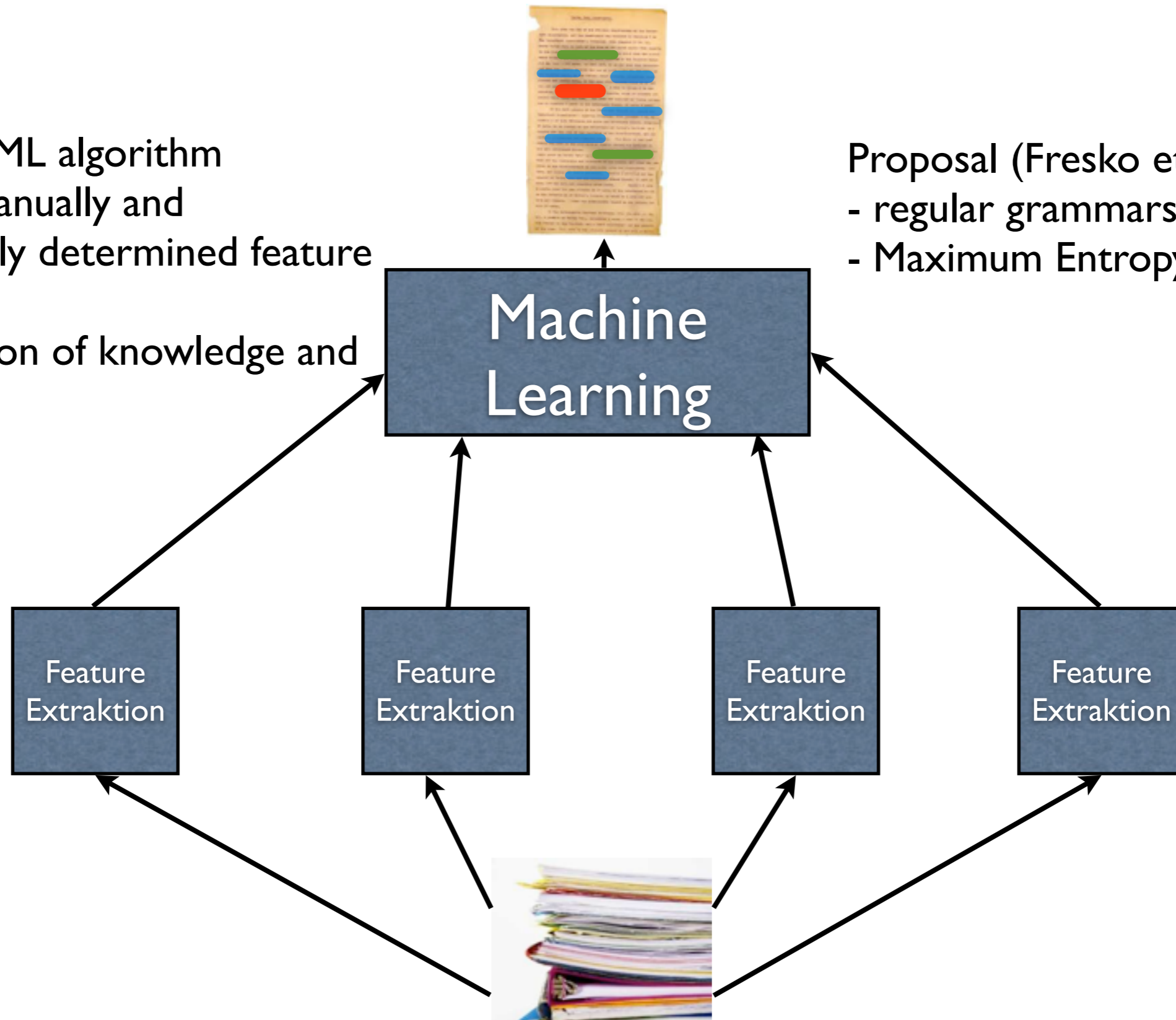


Feature based Strategies

Idea:

- choose a ML algorithm
- choose manually and automatically determined feature templates
- combination of knowledge and statistics

- Proposal (Fresko et al., 2005):
- regular grammars (hand coded)
 - Maximum Entropy Learning

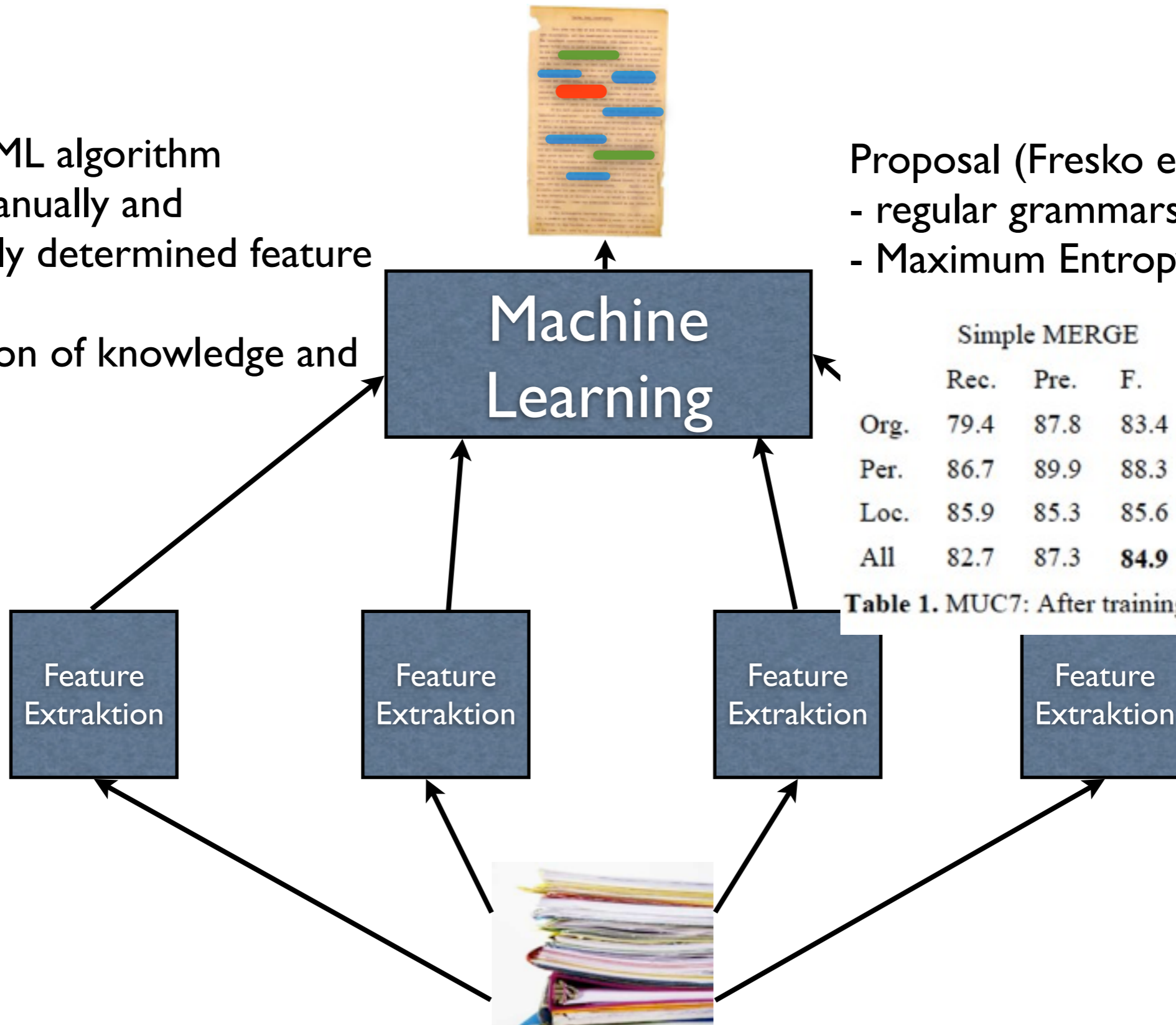


Fresko, Rozenfeld, Feldman „A Hybrid Approach to NER by Integrating Manual Rules into MEM“, CIKM, 2005.

Feature based Strategies

Idea:

- choose a ML algorithm
- choose manually and automatically determined feature templates
- combination of knowledge and statistics



Proposal (Fresko et al., 2005):

- regular grammars (hand coded)
- Maximum Entropy Learning

	Simple MERGE			MERGE+Rules		
	Rec.	Pre.	F.	Rec.	Pre.	F.
Org.	79.4	87.8	83.4	83.9	90.9	87.2
Per.	86.7	89.9	88.3	90.6	93.4	92.0
Loc.	85.9	85.3	85.6	91.5	91.8	91.7
All	82.7	87.3	84.9	87.9	91.7	89.8

Table 1. MUC7: After training with 100 documents.

Fresko, Rozenfeld, Feldman „A Hybrid Approach to NER by Integrating Manual Rules into MEM“, CIKM, 2005.

Co-Training & Bootstrapping

Bootstrapper

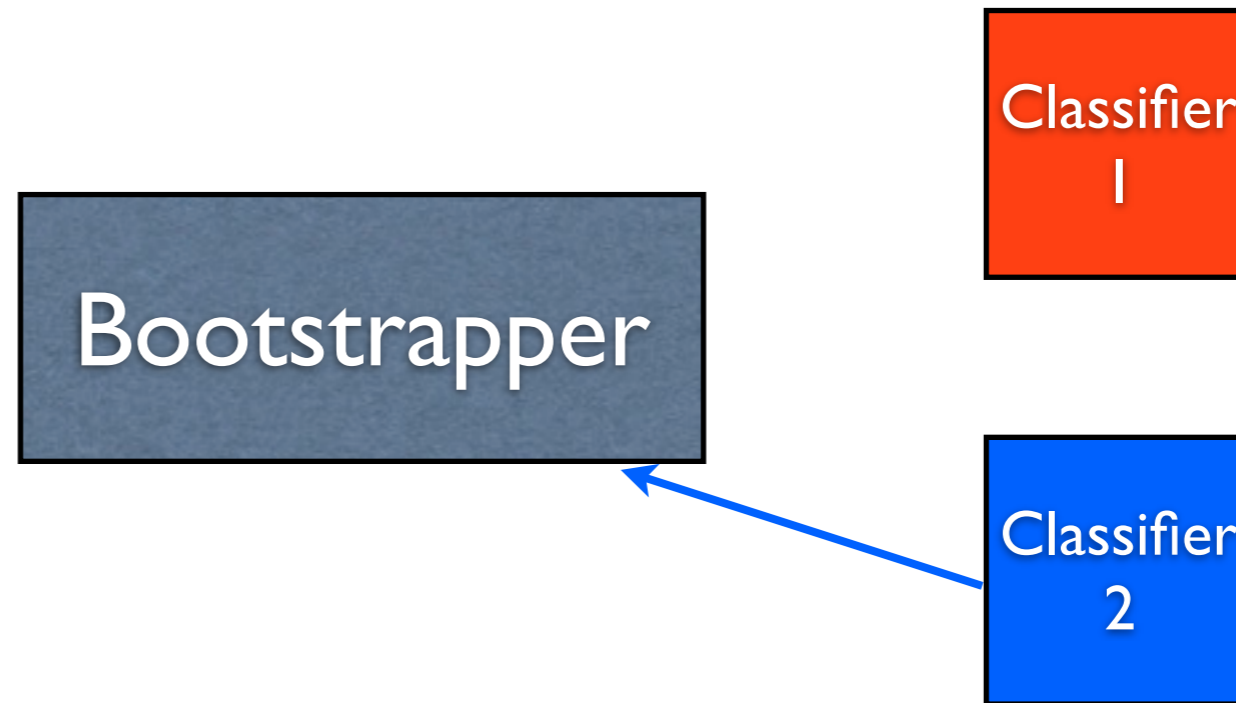
Co-Training & Bootstrapping

Bootstrapper

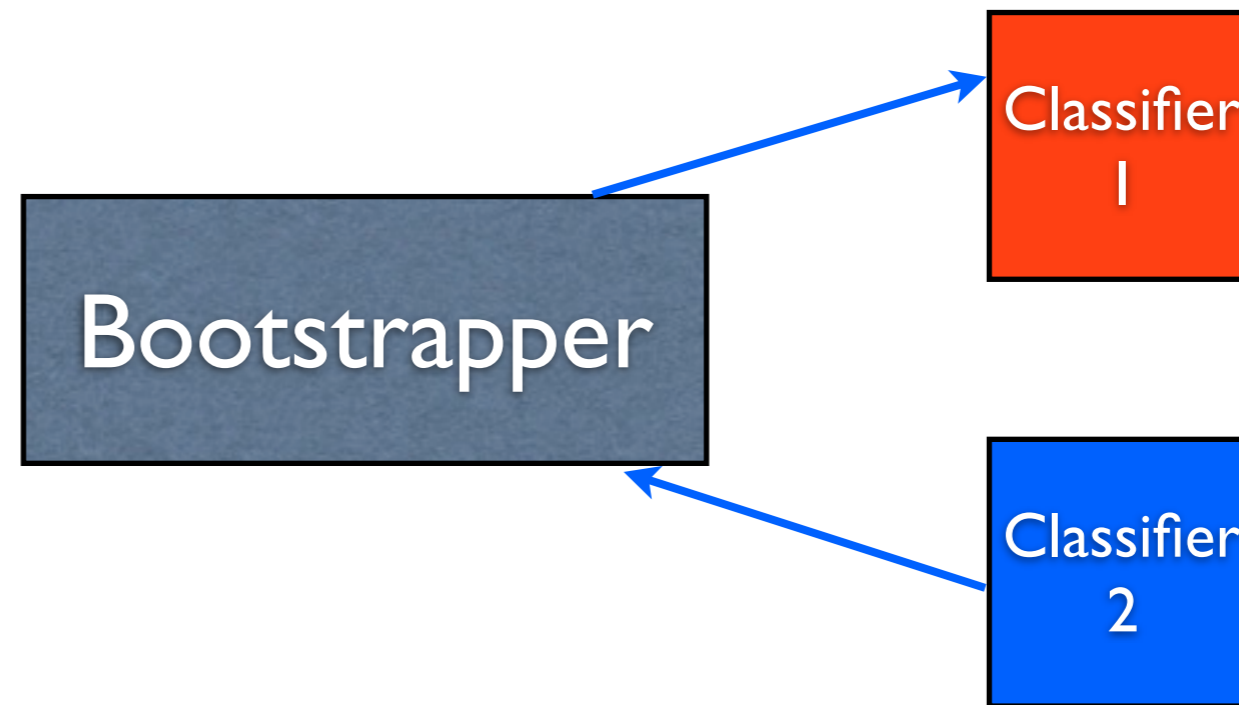
Classifier
1

Classifier
2

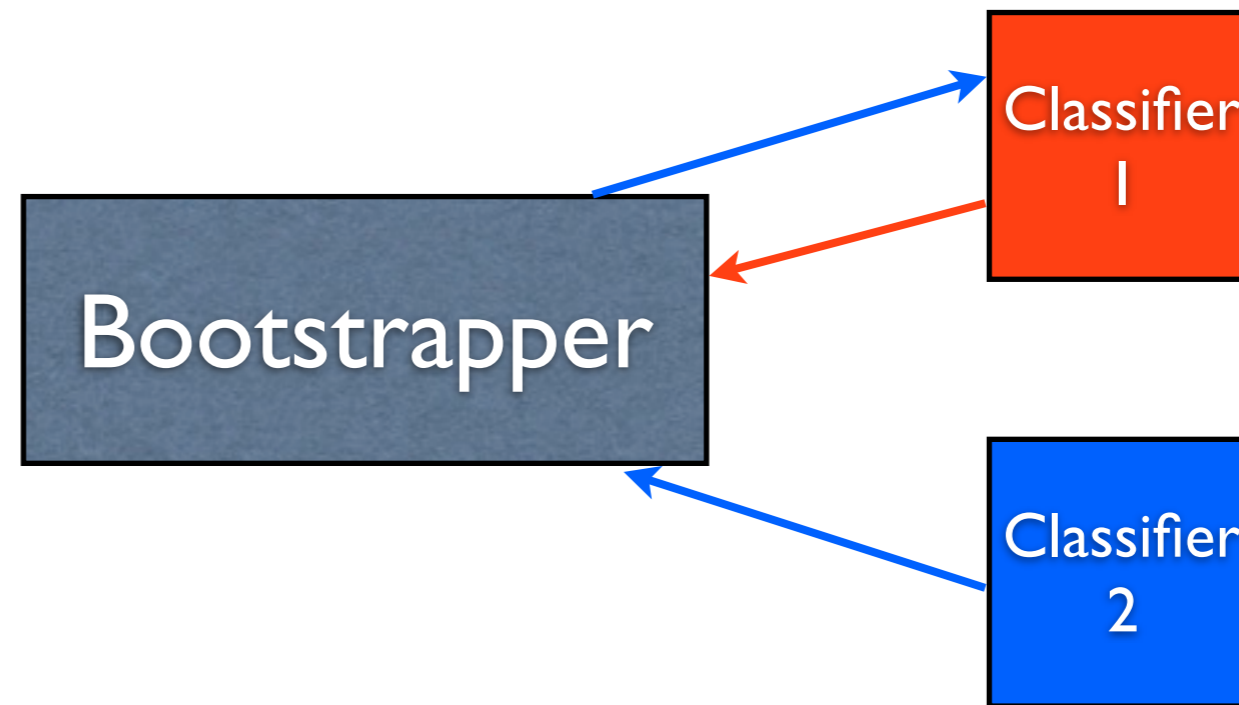
Co-Training & Bootstrapping



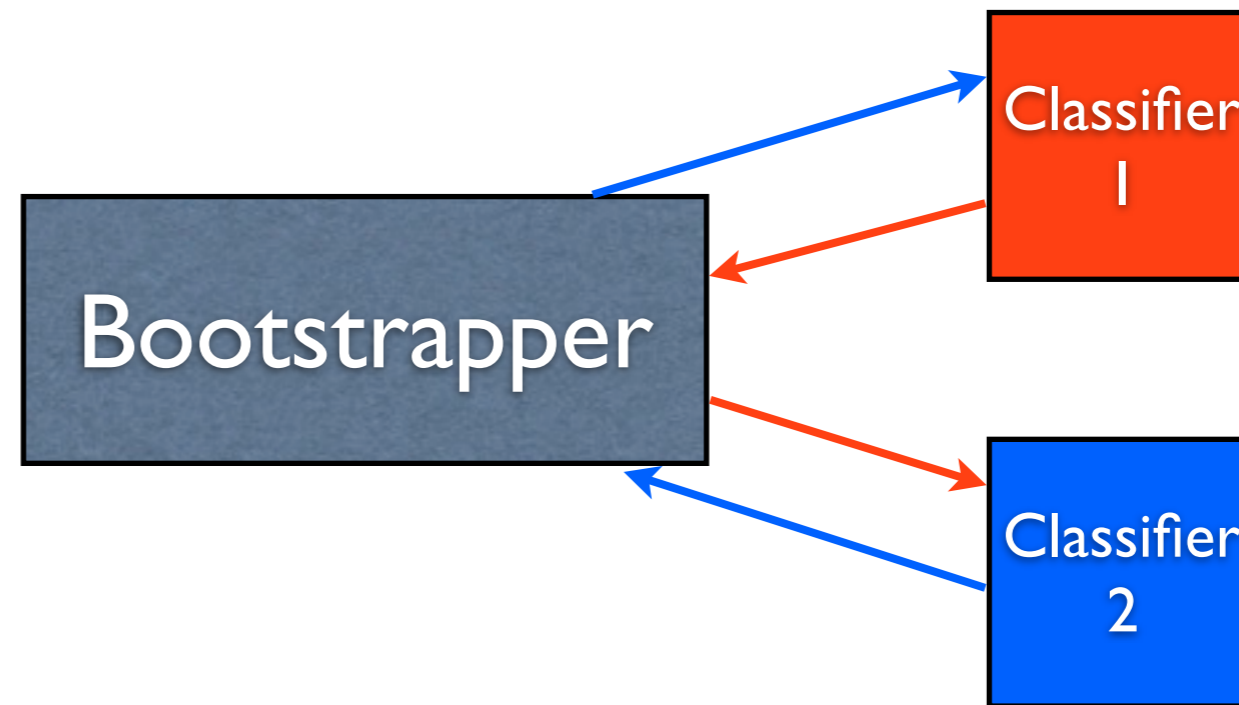
Co-Training & Bootstrapping



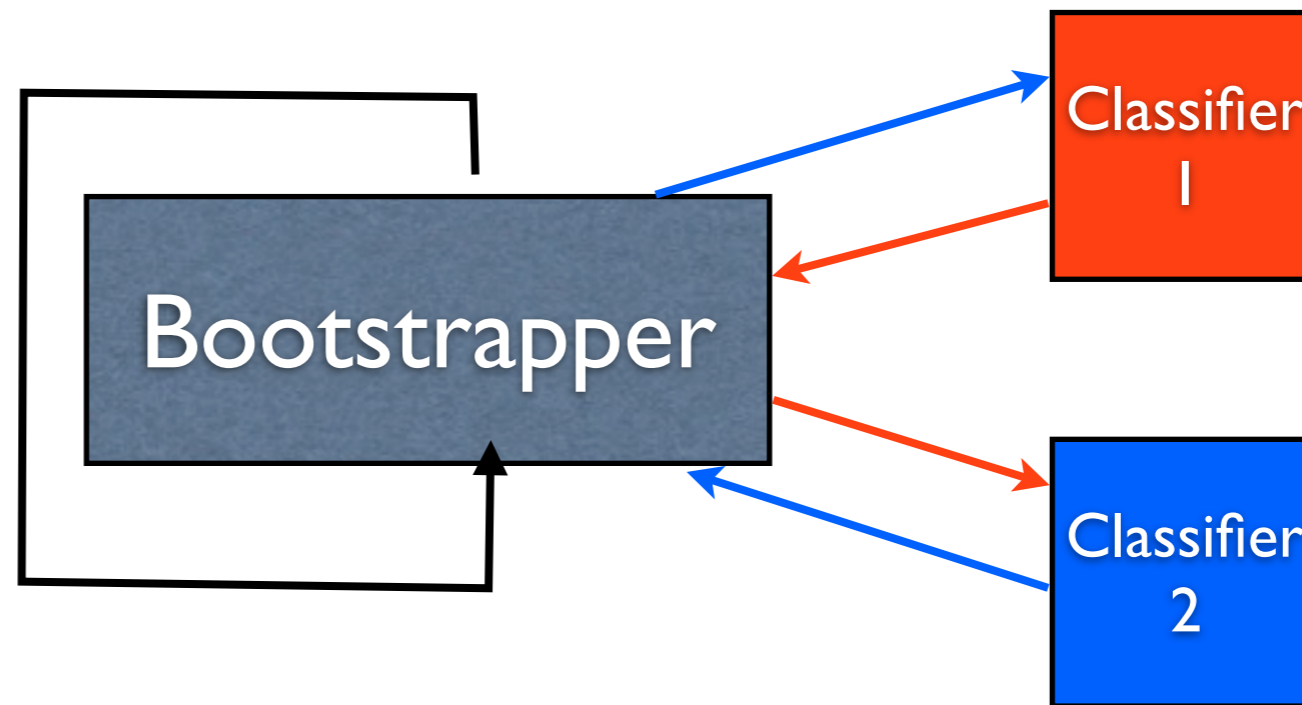
Co-Training & Bootstrapping



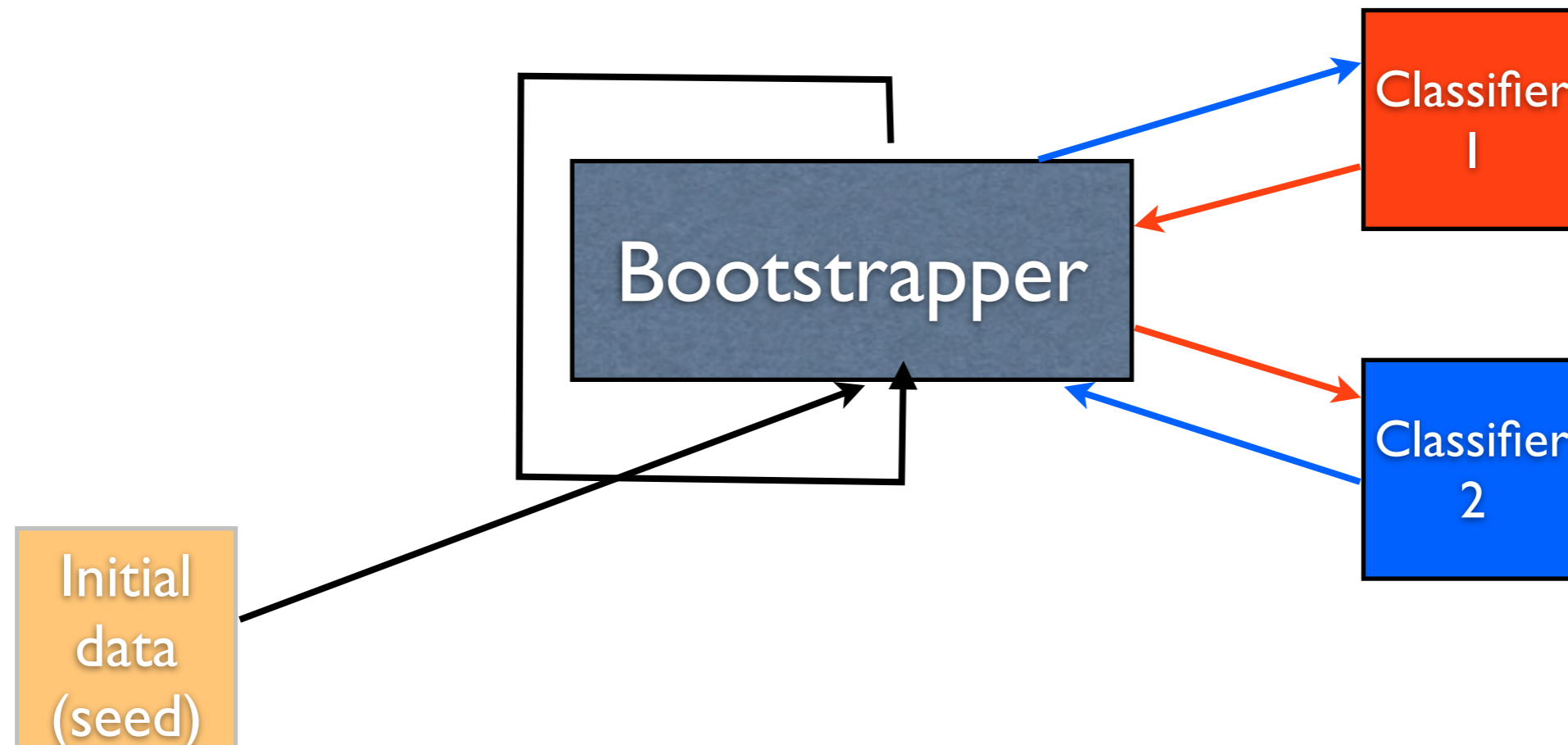
Co-Training & Bootstrapping



Co-Training & Bootstrapping

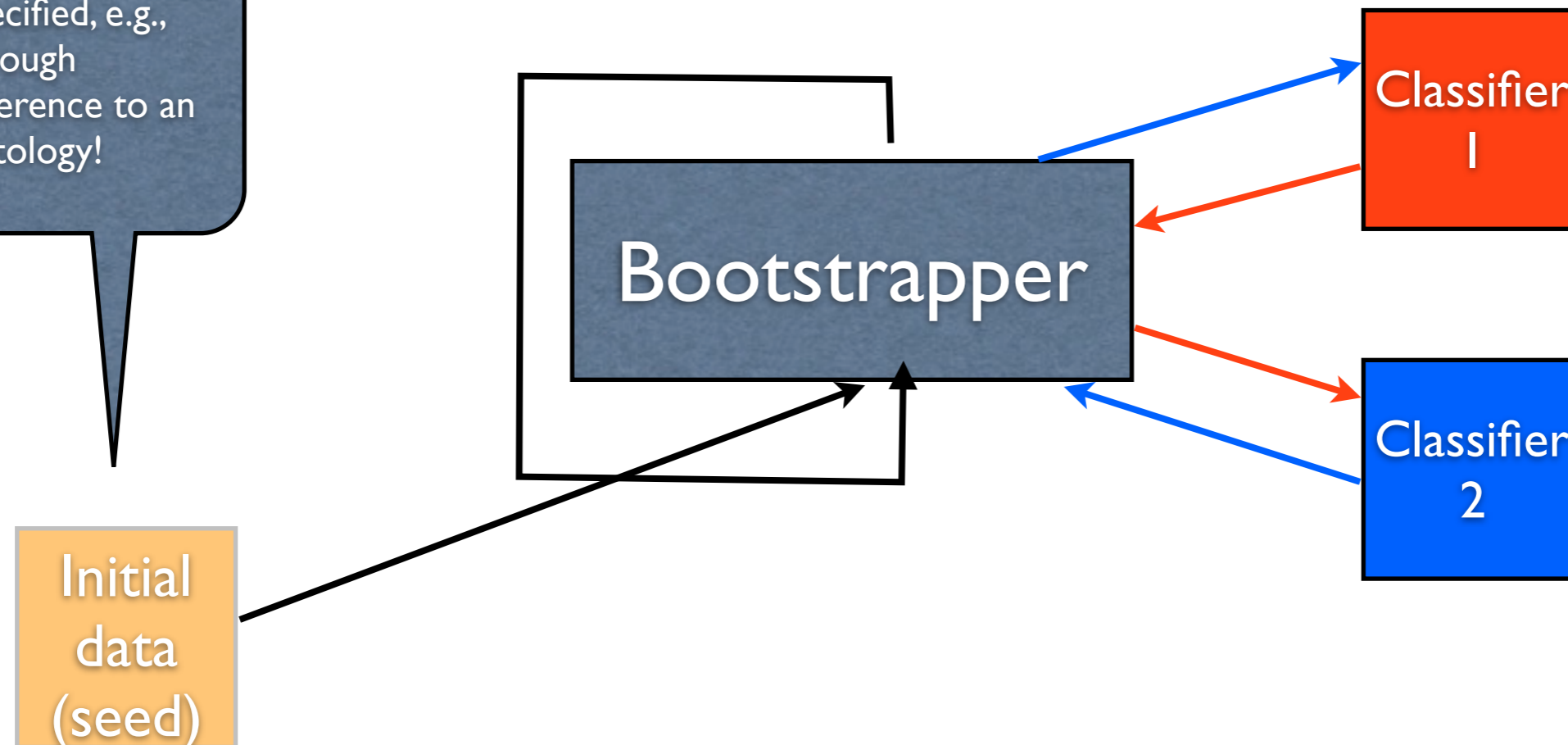


Co-Training & Bootstrapping



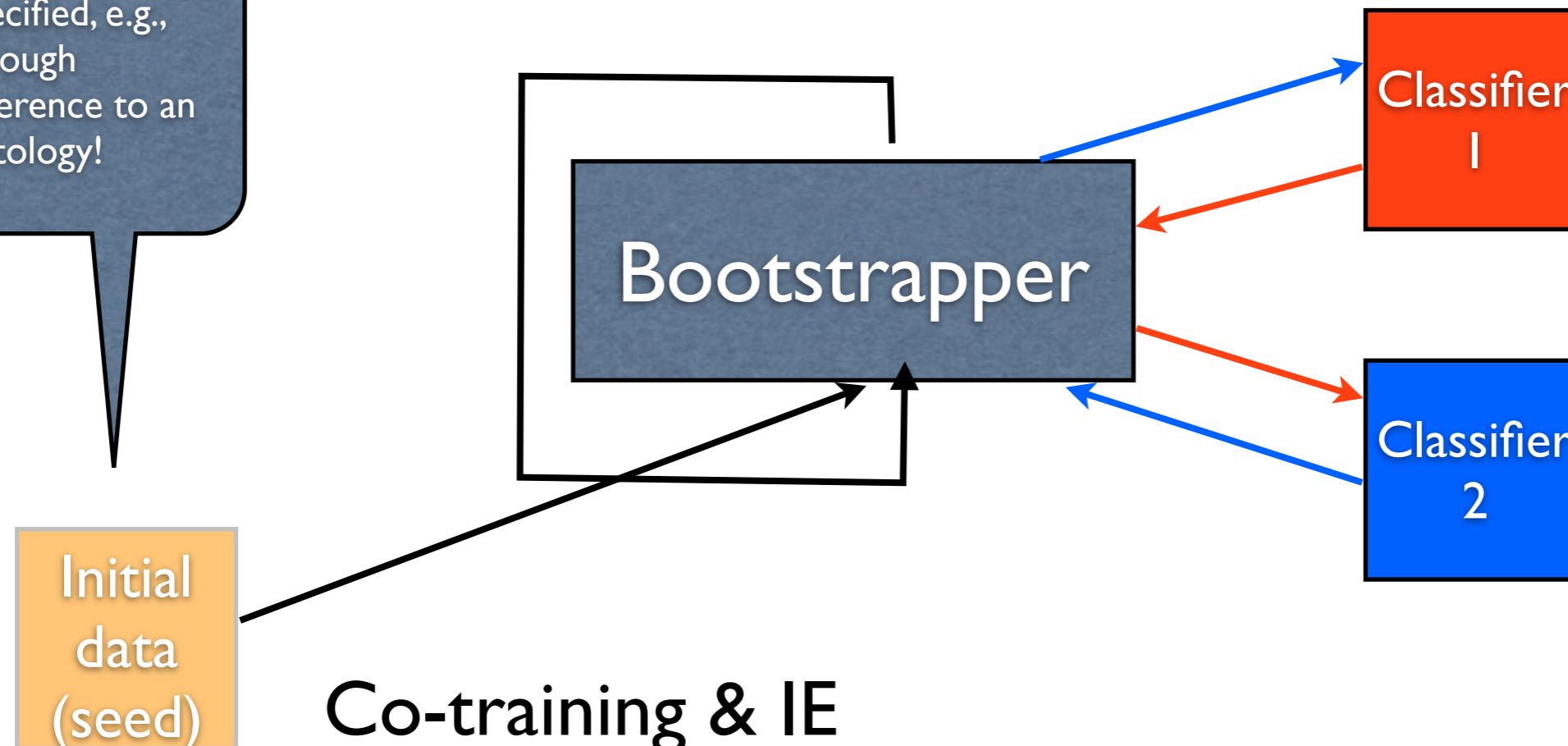
Co-Training & Bootstrapping

Note:
These are manually
specified, e.g.,
through
reference to an
ontology!



Co-Training & Bootstrapping

Note:
These are manually specified, e.g., through reference to an ontology!



Co-training & IE

- NER, cf Singer & Collins, 1999

Interaction of spelling and context features

- REE, cf. Surdeanu et al. 2006

Interaction of text classifier and pattern acquisition

Co-Training & Bootstrapping

Note:
These are
manually
specified, e.g.,

Baseline	Co-training
s(he) o(game)	v(win) o(title)
v(miss) o(game)	s(I) v(play)
v(play) o(game)	s(he) v(game)
v(play) io(in LOC)	s(we) v(play)
v(go) o(be)	v(miss) o(game)
s(he) v(be)	s(he) v(coach)
s(that) v(be)	v(lose) o(game)
s(I) v(be)	s(I) o(play)
s(it) v(go) o(be)	v(make) o(play)
s(it) v(be)	v(play) io(in game)
s(I) v(think)	v(want) o(play)
s(I) v(know)	v(win) o(MISC)
s(I) v(want)	s(he) o(player)
s(there) v(be)	v(start) o(game)
s(we) v(do)	s(PER) o(contract)
v(do) o(it)	s(we) o(play)
s(it) o(be)	s(team) v(win)
s(we) v(are)	v(rush) io(for yard)
s(we) v(go)	s(we) o(team)
s(PER) o(DATE)	v(win) o(Bowl)

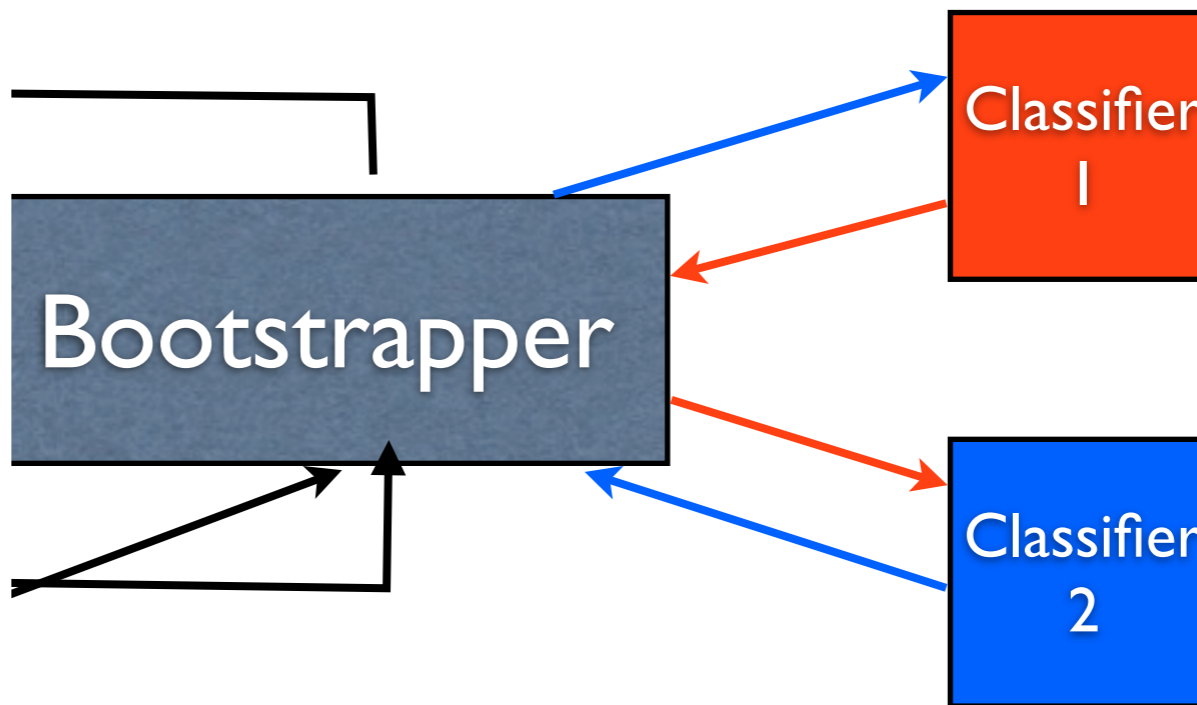


Table 4: Top 20 patterns acquired from the Sports domain by the baseline system (Riloff) and the co-training system for the AP collection. The correct patterns are in bold.

ing & IE

f Singer & Collins, 1999

Interaction of spelling and context features

- REE, cf. Surdeanu et al. 2006

Interaction of text classifier and pattern acquisition

QA and Hybrid IE

- Observation: answer extraction is a kind of question-driven IE (NER and REE)

Where does Bill Gates live? `lives_in(Town:?, Pers:Bill Gates)`

What is a CEO? `is_a(Pos:CEO, Conc:?)`

QA and Hybrid IE

- Observation: answer extraction is a kind of question-driven IE (NER and REE)

Where does Bill Gates live? lives_in(Town:?, Pers:Bill Gates)

What is a CEO? is_a(Pos:CEO, Conc:?)

Domain open answering of definition questions from the Web

QA and Hybrid IE

- Observation: answer extraction is a kind of question-driven IE (NER and REE)

Where does Bill Gates live? lives_in(Town:?, Pers:Bill Gates)

What is a CEO? is_a(Pos:CEO, Conc:?)

Domain open answering of definition questions from the Web



Figuroa, A., Neumann, G. and Atkinson, J. (2009) Searching for Definitional Answers on the Web using Surface Patterns. In journal IEEE Computer volume 42 number 4, Pages 68-76, IEEE, 4/2009.

QA and Hybrid IE

- Observation: answer extraction is a kind of question-driven IE (NER and REE)

Where does Bill Gates live? lives_in(Town:?, Pers:Bill Gates)

What is a CEO? is_a(Pos:CEO, Conc:?)

Domain open answering of definition questions from the Web



Problem:
How to find optimal ranking of answer candidates?

Figueroa, A., Neumann, G. and Atkinson, J. (2009) Searching for Definitional Answers on the Web using Surface Patterns. In journal IEEE Computer volume 42 number 4, Pages 68-76, IEEE, 4/2009.

Wikipedia as Blueprint!

- Learn from Wikipedia, what a good verbalization of a definition looks like!



Wikipedia as Blueprint!

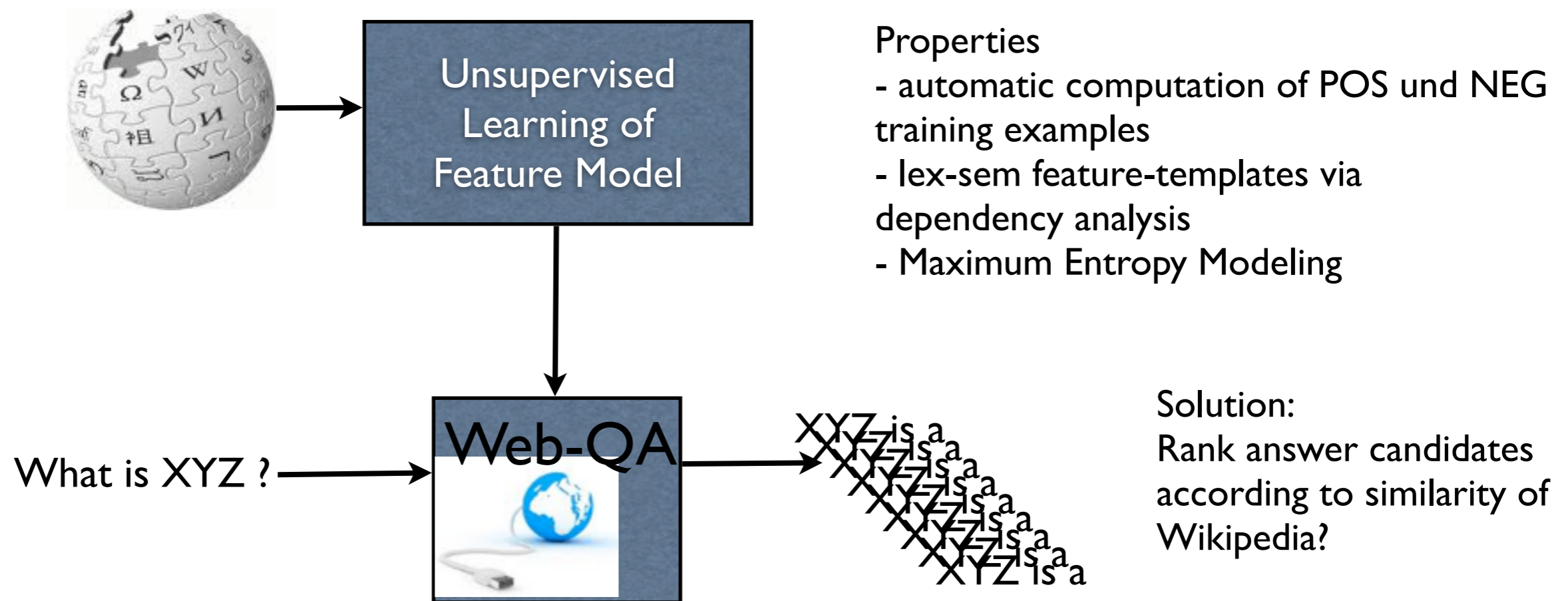
- Learn from Wikipedia, what a good verbalization of a definition looks like!



Solution:
Rank answer candidates
according to similarity of
Wikipedia?

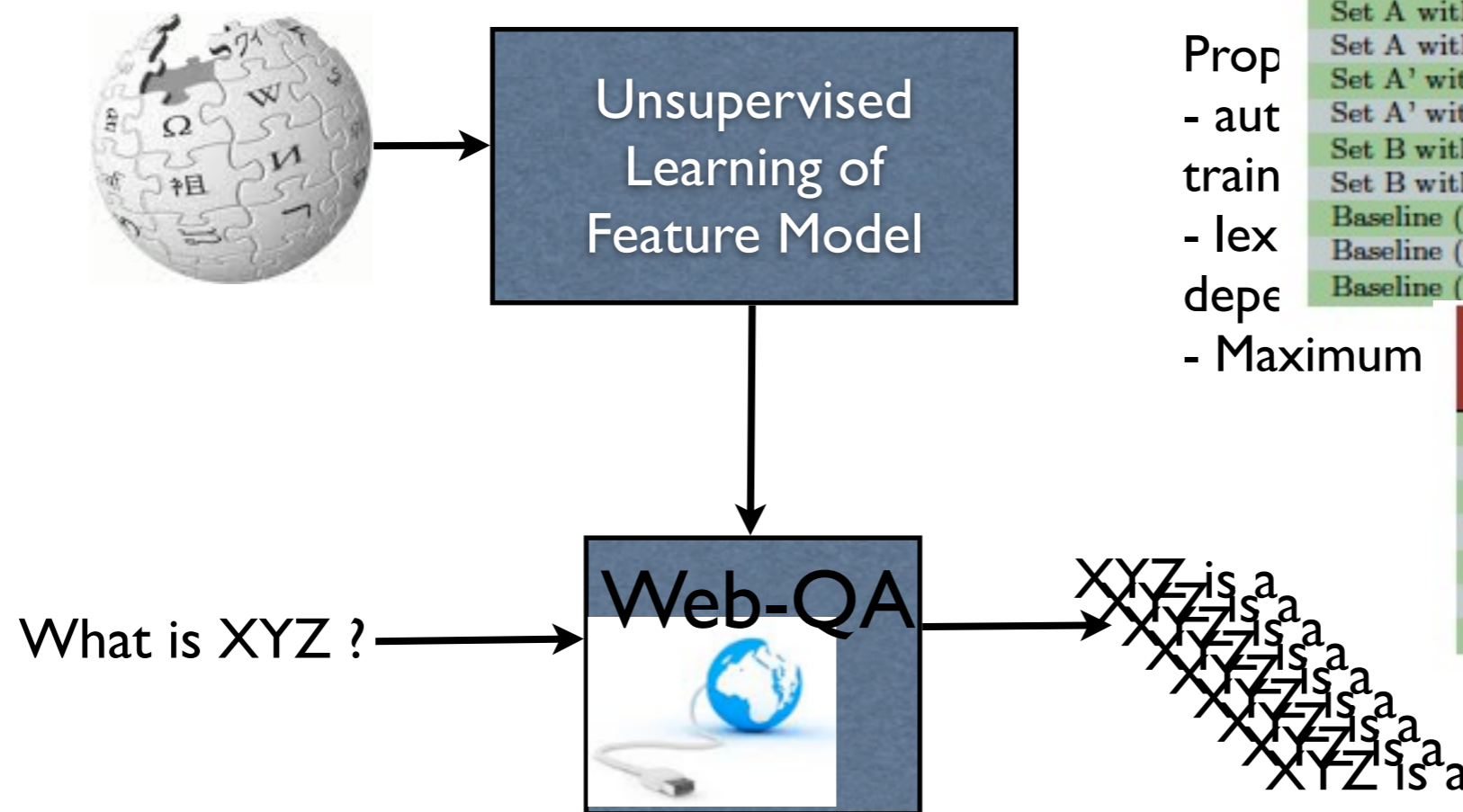
Wikipedia as Blueprint!

- Learn from Wikipedia, what a good verbalization of a definition looks like!



Wikipedia as Blueprint!

- Learn from Wikipedia, what a good verbalization of a definition looks like!



Prop
- aut
train
- lex
depe
- Maximum

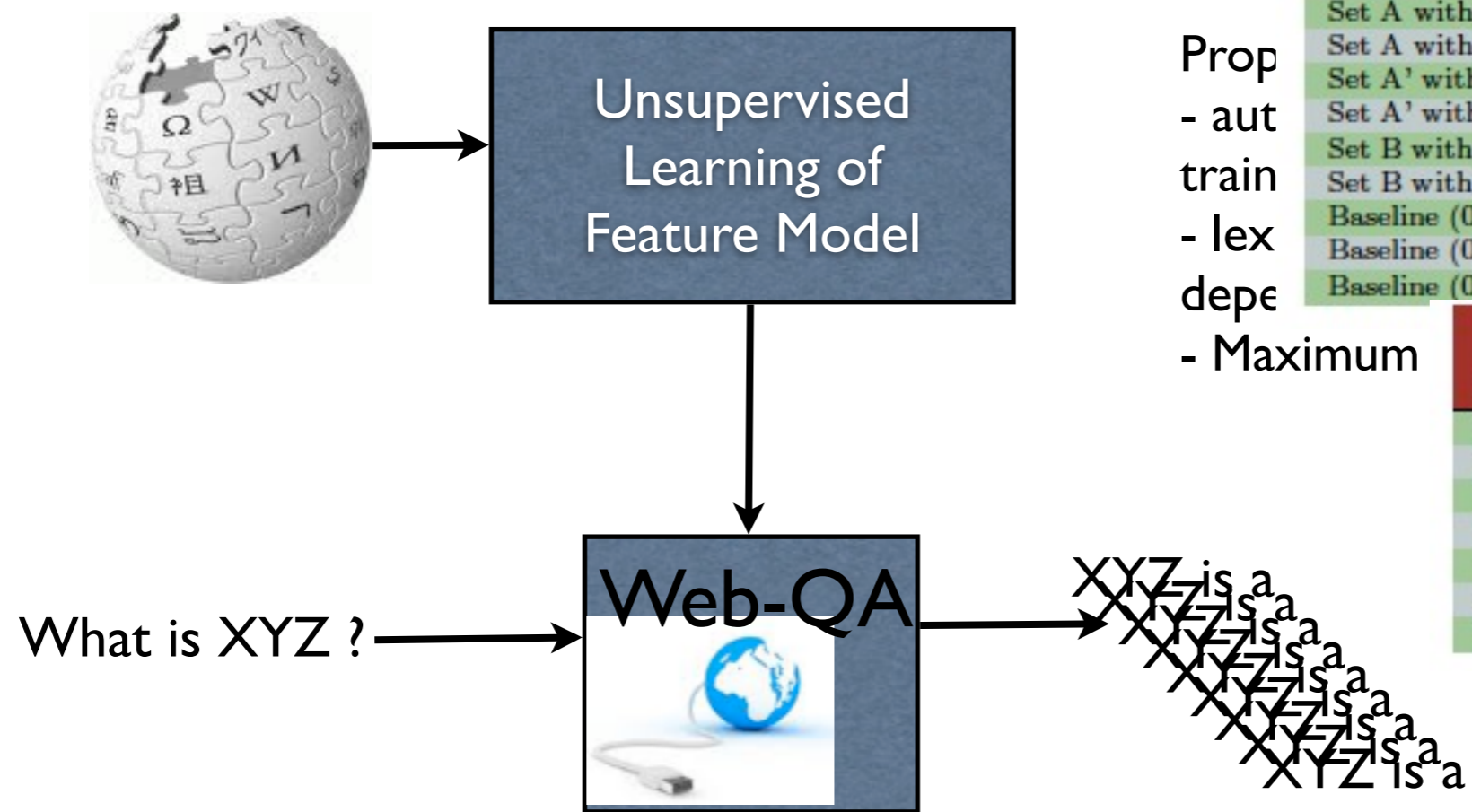
Best Attributes Found for	Applied to	
	Set A'	Set A
	Accuracy	Accuracy
Set A without NLP	81.16%	
Set A with NLP	85.94%	
Set A' without NLP		78.28%
Set A' with NLP		83.04%
Set B without NLP	76.86%	75.50%
Set B with NLP	63.19%	74.61%
Baseline (0.3)	54.13%	44.17%
Baseline (0.2)	54.63%	44.23%
Baseline (0.1)	56.25%	44.77%

Best Attributes Found for	Applied to
	Set B
	Accuracy
Set A without NLP	59.86%
Set A with NLP	58.71%
Set A' without NLP	58.25%
Set A' with NLP	58.44%
Baseline (0.3)	56.26%
Baseline (0.2)	57.77%
Baseline (0.1)	56.73%

according to similarity of Wikipedia?

Wikipedia as Blueprint!

- Learn from Wikipedia, what a good verbalization of a definition looks like!



Prop
- aut
train
- lex
depe
- Maximum

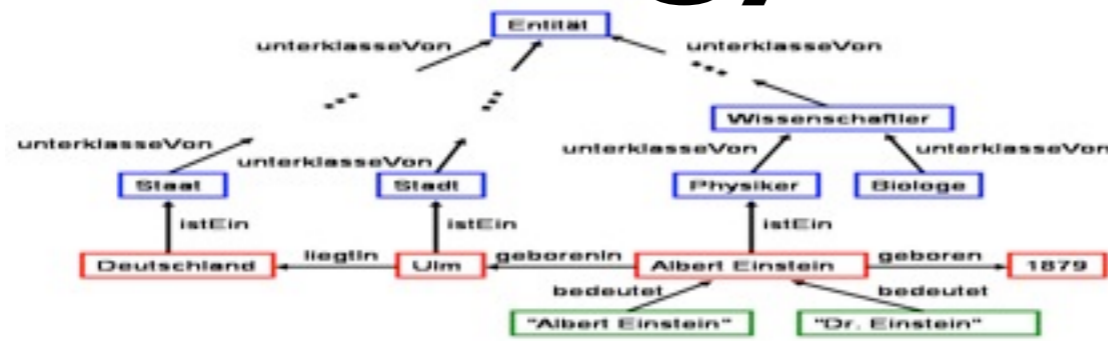
Best Attributes Found for	Applied to	
	Set A'	Set A
	Accuracy	Accuracy
Set A without NLP	81.16%	
Set A with NLP	85.94%	
Set A' without NLP		78.28%
Set A' with NLP		83.04%
Set B without NLP	76.86%	75.50%
Set B with NLP	63.19%	74.61%
Baseline (0.3)	54.13%	44.17%
Baseline (0.2)	54.63%	44.23%
Baseline (0.1)	56.25%	44.77%

Best Attributes Found for	Applied to
	Set B
	Accuracy
Set A without NLP	59.86%
Set A with NLP	58.71%
Set A' without NLP	58.25%
Set A' with NLP	58.44%
Baseline (0.3)	56.26%
Baseline (0.2)	57.77%
Baseline (0.1)	56.73%

according to similarity of Wikipedia?

Remark: Method is a step towards Web-scalable ontology learning.

Ontology based IE



IE

Computer System Name	City	Year
Computer System 001	Frankfurt	1950
Computer System 002	Frankfurt	1951
Computer System 003	Frankfurt	1952
Computer System 004	Frankfurt	1953
Computer System 005	Frankfurt	1954
Computer System 006	Frankfurt	1955
Computer System 007	Frankfurt	1956
Computer System 008	Frankfurt	1957
Computer System 009	Frankfurt	1958
Computer System 010	Frankfurt	1959
Computer System 011	Frankfurt	1960
Computer System 012	Frankfurt	1961
Computer System 013	Frankfurt	1962
Computer System 014	Frankfurt	1963
Computer System 015	Frankfurt	1964
Computer System 016	Frankfurt	1965
Computer System 017	Frankfurt	1966
Computer System 018	Frankfurt	1967
Computer System 019	Frankfurt	1968
Computer System 020	Frankfurt	1969
Computer System 021	Frankfurt	1970
Computer System 022	Frankfurt	1971
Computer System 023	Frankfurt	1972
Computer System 024	Frankfurt	1973
Computer System 025	Frankfurt	1974
Computer System 026	Frankfurt	1975
Computer System 027	Frankfurt	1976
Computer System 028	Frankfurt	1977
Computer System 029	Frankfurt	1978
Computer System 030	Frankfurt	1979
Computer System 031	Frankfurt	1980
Computer System 032	Frankfurt	1981
Computer System 033	Frankfurt	1982
Computer System 034	Frankfurt	1983
Computer System 035	Frankfurt	1984
Computer System 036	Frankfurt	1985
Computer System 037	Frankfurt	1986
Computer System 038	Frankfurt	1987
Computer System 039	Frankfurt	1988
Computer System 040	Frankfurt	1989
Computer System 041	Frankfurt	1990
Computer System 042	Frankfurt	1991
Computer System 043	Frankfurt	1992
Computer System 044	Frankfurt	1993
Computer System 045	Frankfurt	1994
Computer System 046	Frankfurt	1995
Computer System 047	Frankfurt	1996
Computer System 048	Frankfurt	1997
Computer System 049	Frankfurt	1998
Computer System 050	Frankfurt	1999
Computer System 051	Frankfurt	2000
Computer System 052	Frankfurt	2001
Computer System 053	Frankfurt	2002
Computer System 054	Frankfurt	2003
Computer System 055	Frankfurt	2004
Computer System 056	Frankfurt	2005
Computer System 057	Frankfurt	2006
Computer System 058	Frankfurt	2007
Computer System 059	Frankfurt	2008
Computer System 060	Frankfurt	2009
Computer System 061	Frankfurt	2010
Computer System 062	Frankfurt	2011
Computer System 063	Frankfurt	2012
Computer System 064	Frankfurt	2013
Computer System 065	Frankfurt	2014
Computer System 066	Frankfurt	2015
Computer System 067	Frankfurt	2016
Computer System 068	Frankfurt	2017
Computer System 069	Frankfurt	2018
Computer System 070	Frankfurt	2019
Computer System 071	Frankfurt	2020
Computer System 072	Frankfurt	2021
Computer System 073	Frankfurt	2022
Computer System 074	Frankfurt	2023
Computer System 075	Frankfurt	2024
Computer System 076	Frankfurt	2025
Computer System 077	Frankfurt	2026
Computer System 078	Frankfurt	2027
Computer System 079	Frankfurt	2028
Computer System 080	Frankfurt	2029
Computer System 081	Frankfurt	2030

Ontology based IE



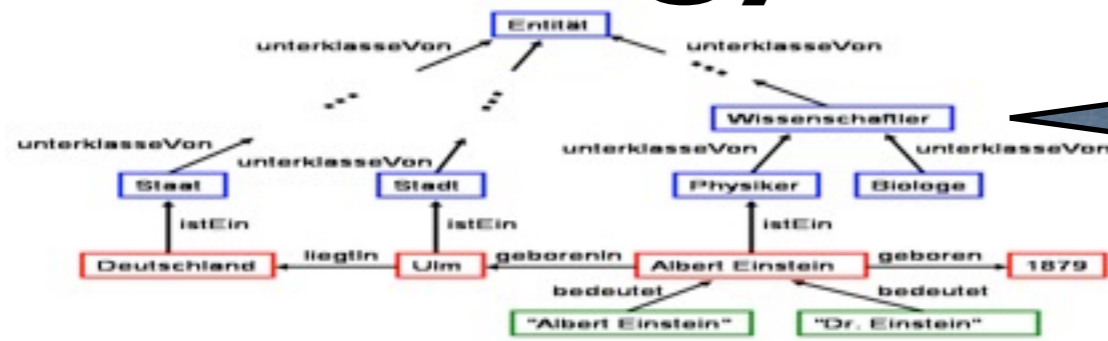
The ontology defines the type of the information, which has to be extracted from texts: e.g., types of or institutions and their inter relationship. It defines the structure of the data base, which has to be extracted automatically with the help of OBIE.



IE

Computer System Name	Manufacturer	Model
Computer System 001	IBM	Model 308
Computer System 002	IBM	Model 309
Computer System 003	IBM	Model 310
Computer System 004	IBM	Model 311
Computer System 005	IBM	Model 312
Computer System 006	IBM	Model 313
Computer System 007	IBM	Model 314
Computer System 008	IBM	Model 315
Computer System 009	IBM	Model 316
Computer System 010	IBM	Model 317
Computer System 011	IBM	Model 318
Computer System 012	IBM	Model 319
Computer System 013	IBM	Model 320
Computer System 014	IBM	Model 321
Computer System 015	IBM	Model 322
Computer System 016	IBM	Model 323
Computer System 017	IBM	Model 324
Computer System 018	IBM	Model 325
Computer System 019	IBM	Model 326
Computer System 020	IBM	Model 327
Computer System 021	IBM	Model 328
Computer System 022	IBM	Model 329
Computer System 023	IBM	Model 330
Computer System 024	IBM	Model 331
Computer System 025	IBM	Model 332
Computer System 026	IBM	Model 333
Computer System 027	IBM	Model 334
Computer System 028	IBM	Model 335
Computer System 029	IBM	Model 336
Computer System 030	IBM	Model 337
Computer System 031	IBM	Model 338
Computer System 032	IBM	Model 339
Computer System 033	IBM	Model 340
Computer System 034	IBM	Model 341
Computer System 035	IBM	Model 342
Computer System 036	IBM	Model 343
Computer System 037	IBM	Model 344
Computer System 038	IBM	Model 345
Computer System 039	IBM	Model 346
Computer System 040	IBM	Model 347
Computer System 041	IBM	Model 348
Computer System 042	IBM	Model 349
Computer System 043	IBM	Model 350
Computer System 044	IBM	Model 351
Computer System 045	IBM	Model 352
Computer System 046	IBM	Model 353
Computer System 047	IBM	Model 354
Computer System 048	IBM	Model 355
Computer System 049	IBM	Model 356
Computer System 050	IBM	Model 357

Ontology based IE



The ontology defines the type of the information, which has to be extracted from texts: e.g., types of or institutions and their inter relationship. It defines the structure of the data base, which has to be extracted automatically with the help of OBIE.



IE

Ontology population
ontology learning

Computer System Name	Manufacturer	Model
Computer System 0001	IBM	Model 3084
Computer System 0002	IBM	Model 3084
Computer System 0003	IBM	Model 3084
Computer System 0004	IBM	Model 3084
Computer System 0005	IBM	Model 3084
Computer System 0006	IBM	Model 3084
Computer System 0007	IBM	Model 3084
Computer System 0008	IBM	Model 3084
Computer System 0009	IBM	Model 3084
Computer System 0010	IBM	Model 3084
Computer System 0011	IBM	Model 3084
Computer System 0012	IBM	Model 3084
Computer System 0013	IBM	Model 3084
Computer System 0014	IBM	Model 3084
Computer System 0015	IBM	Model 3084
Computer System 0016	IBM	Model 3084
Computer System 0017	IBM	Model 3084
Computer System 0018	IBM	Model 3084
Computer System 0019	IBM	Model 3084
Computer System 0020	IBM	Model 3084

Ontology based IE



The ontology defines the type of the information, which has to be extracted from texts: e.g., types of or institutions and their inter relationship. It defines the structure of the data base, which has to be extracted automatically with the help of OBIE.

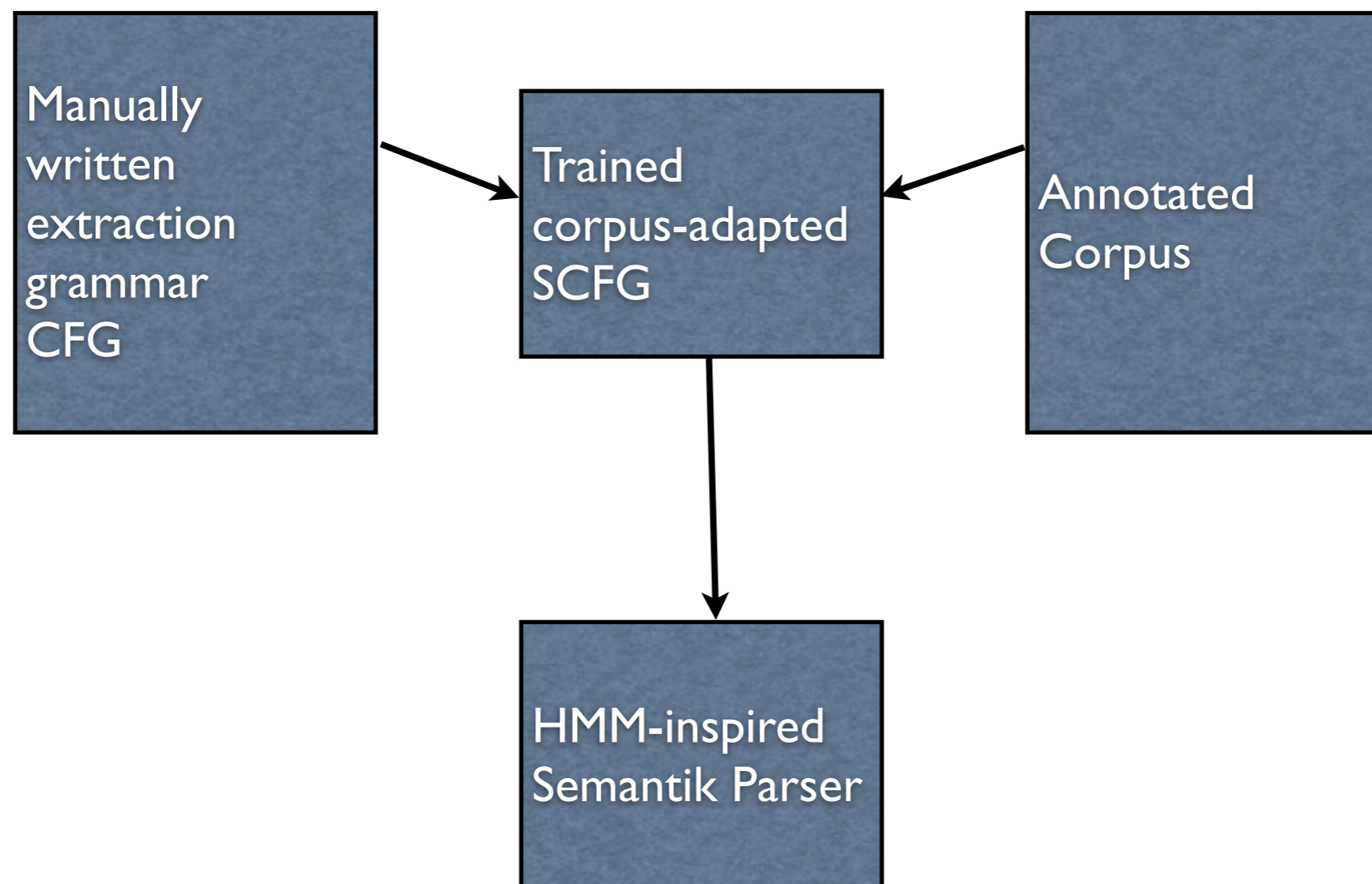


IE

Computer System Name	Manufacturer	Model
Computer System 001	IBM	Model 308
Computer System 002	IBM	Model 308
Computer System 003	IBM	Model 308
Computer System 004	IBM	Model 308
Computer System 005	IBM	Model 308
Computer System 006	IBM	Model 308
Computer System 007	IBM	Model 308
Computer System 008	IBM	Model 308
Computer System 009	IBM	Model 308
Computer System 010	IBM	Model 308
Computer System 011	IBM	Model 308
Computer System 012	IBM	Model 308
Computer System 013	IBM	Model 308
Computer System 014	IBM	Model 308
Computer System 015	IBM	Model 308
Computer System 016	IBM	Model 308
Computer System 017	IBM	Model 308
Computer System 018	IBM	Model 308
Computer System 019	IBM	Model 308
Computer System 020	IBM	Model 308

Ontology population
ontology learning

TEG - Tree Extraction Grammars



Rosenfeld, Feldman & Freski „TEG - a hybrid approach to information extraction“, Knowledge Information Systems (2006) 1-18.

TEG - Example

Hand coded grammars

nonterm start Text;

concept Person;

ngram NGFirstName;

ngram NGLastName;

ngram NGNone;

termlist TLHonorific = Mr Mrs Miss Ms Dr;

(1) Person :- TLHonorific NGLastName;

(2) Person :- NGFirstName NGLastName;

(3) Text :- NGNone Text;

(4) Text :- Person Text;

(5) Text :- ;

TEG - Example

Hand coded grammars

nonterm start Text;

concept Person;

ngram NGFirstName;

ngram NGLastName;

ngram NGNone;

termlist TLHonorific = Mr Mrs Miss Ms Dr;

(1) Person :- TLHonorific NGLastName;

(2) Person :- NGFirstName NGLastName;

(3) Text :- NGNone Text;

(4) Text :- Person Text;

(5) Text :- ;

Yesterday, <Person> Dr Simmons </Person>, the distinguished scientist presented the discovery.

TEG - Example

Hand coded grammars

nonterm start Text;

concept Person;

ngram NGFirstName;

ngram NGLastName;

ngram NGNone;

termlist TLHonorific = Mr Mrs Miss Ms Dr;

(1) Person :- TLHonorific NGLastName;

(2) Person :- NGFirstName NGLastName;

(3) Text :- NGNone Text;

(4) Text :- Person Text;

(5) Text :- ;

Parse corpus



Yesterday, <Person> Dr Simmons </Person>, the distinguished scientist presented the discovery.

TEG - Example

Hand coded grammars

nonterm start Text;

concept Person;

ngram NGFirstName;

ngram NGLastName;

ngram NGNone;

termlist TLHonorific = Mr Mrs Miss Ms Dr;

(1) Person :- TLHonorific NGLastName;

(2) Person :- NGFirstName NGLastName;

(3) Text :- NGNone Text;

(4) Text :- Person Text;

(5) Text :- ;

Parse corpus

Collect statistics

$P(\text{Dr} \mid \text{TLHonorific}) = 1/5$ (choice of one term among five equiprobable ones),

$P(\text{Dr} \mid \text{NGFirstName}) \approx 1/N$, where N is the number of all known words (untrained ngram behaviour).

Yesterday, <Person> Dr Simmons </Person>, the distinguished scientist presented the discovery.

TEG - Example

Hand coded grammars

nonterm start Text;

concept Person;

ngram NGFirstName;

ngram NGLastName;

ngram NGNone;

termlist TLHonorific = Mr Mrs Miss Ms Dr;

(1) Person :- TLHonorific NGLastName;

(2) Person :- NGFirstName NGLastName;

(3) Text :- NGNone Text;

(4) Text :- Person Text;

(5) Text :- ;

Parse corpus

Collect statistics

Yesterday, <Person> Dr Simmons </Person>, the distinguished scientist presented the discovery.

termlist TLHonorific = Mr Mrs Miss Ms <2>Dr;

Person :- <2>TLHonorific NGLastName;

Text :- <11>NGNone Text;

Text :- <2>Person Text;

Text :- <2>;

adapt rules

$P(\text{Dr} \mid \text{TLHonorific}) = 1/5$ (choice of one term among five equiprobable ones),

$P(\text{Dr} \mid \text{NGFirstName}) \approx 1/N$, where N is the number of all known words (untrained ngram behaviour).

TEG - Experiments

MUC-7 NER task

	HMM entity extractor			Emulation using TEG			DIAL Rules			Full TEG system		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
Person	86.91	85.13	86.01	86.31	86.83	86.57	81.32	93.75	87.53	93.75	90.78	92.24
Org	87.94	89.75	88.84	85.94	89.53	87.7	82.74	93.36	88.05	89.49	90.9	90.19
Location	86.12	87.2	86.66	83.93	90.12	86.91	91.46	89.53	90.49	87.05	94.42	90.58

ACE-2 relation extraction

	HMM entity extractor			Markovian SCFG			Full TEG system (with 7 ROLE rules)		
	Recall	Prec	F	Recall	Prec	F	Recall	Prec	F
Role				67.55	69.86	68.69	83.44	77.3	80.25
Person	85.54	83.22	84.37	89.19	80.19	84.45	89.82	81.68	85.56
Organization	52.62	64.735	58.05	53.57	67.46	59.71	59.49	71.06	64.76
GPE	85.54	83.22	84.37	86.74	84.96	85.84	88.83	84.94	86.84

INC relation extraction

	Partial match results			Exact match results		
	Recall	Prec	F	Recall	Prec	F
PersonAffiliation	89.61	94.52	92.00	75.33	79.46	77.33
OrgLocation	85.32	77.78	80.00	76.47	72.22	74.29
Acquisition	76.00	86.36	80.85	68.00	77.27	72.34

TEG - Potential

- Advantages
 - precise rules can be specified for arbitrary IE applications
 - external knowledge sources can be integrated via termlist
 - ngram-context for terminals via ngram (usable for disambiguation)
 - external systems can be integrated
 - „ngram ngOrgNoun featureset ExtPoS restriction Noun;“
- Possible innovations
 - Constraint based formalism as basis for grammar
 - Specialized parsing algorithms (e.g., supertagging)
 - Ontologies as basis for termlist
 - Extending grammars on basis of bootstrapping (human-controlled)
 - ...

Conclusion

- Hybrid IE as innovative plus for IE research and development.
- There exists already a number of promising and exciting approaches.
- High innovation potential to bring language technology, knowledge-based and statistical system under one umbrella.
- E.g., Multilingual Information Extraction
- E.g., Multi-Channel Information Extraction