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

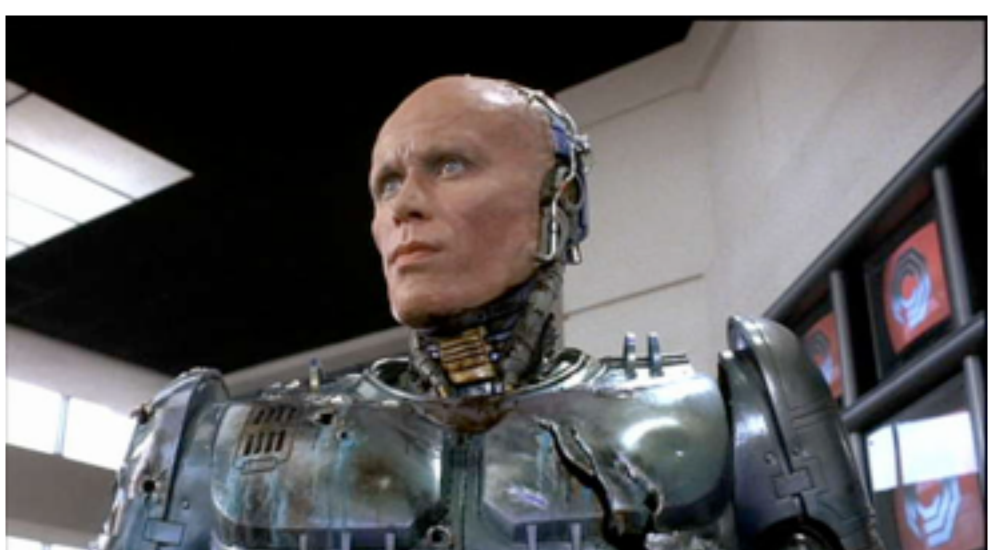
Examples

hybrid engine

Examples

hybrid engine

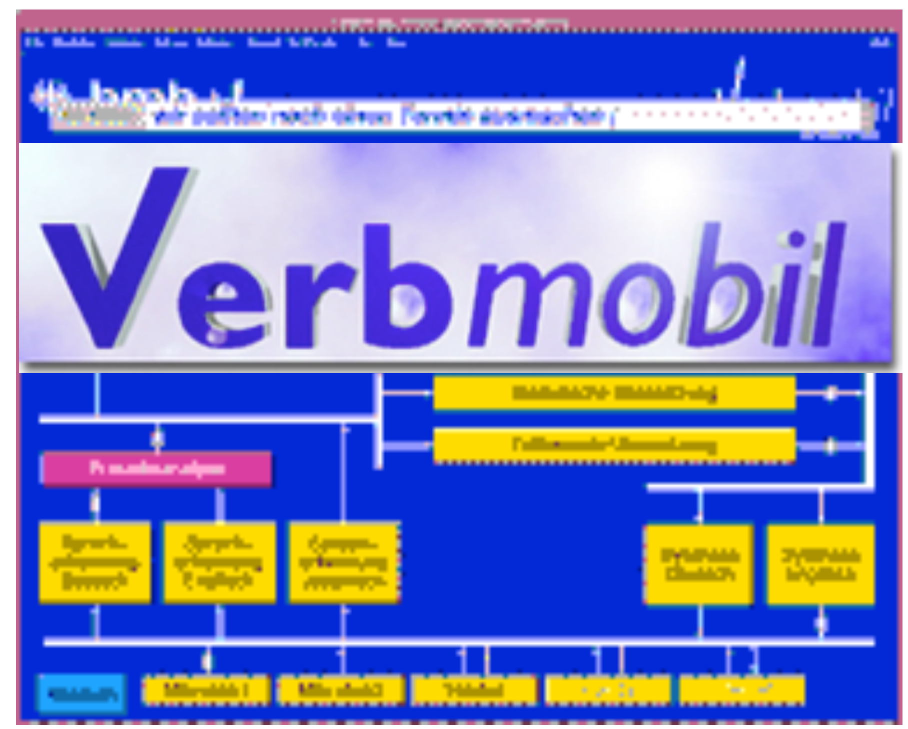
HumanMachine



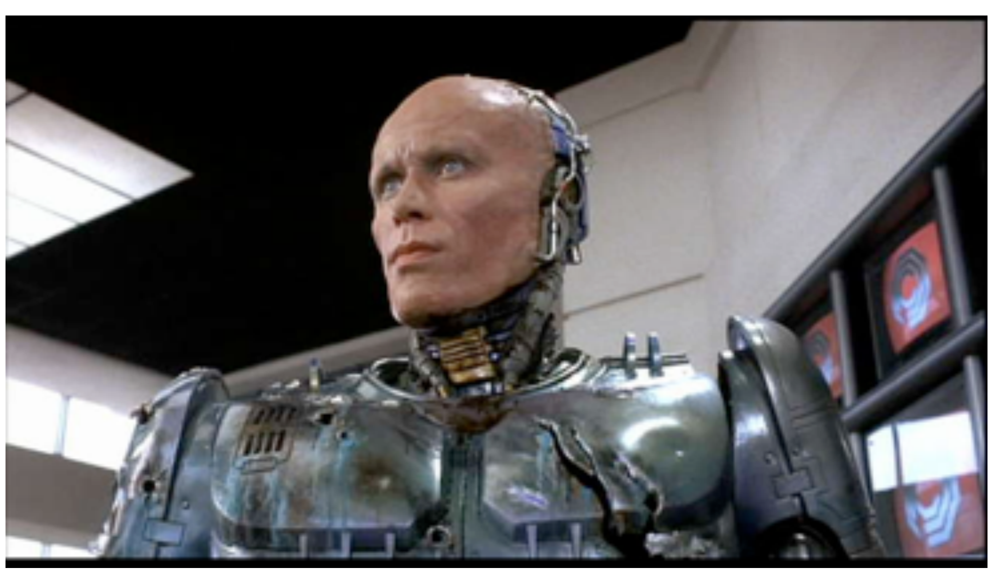
Examples

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

Microsoft
is a Company

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

Seattle is a Location

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

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

Bill Gates
is a Person

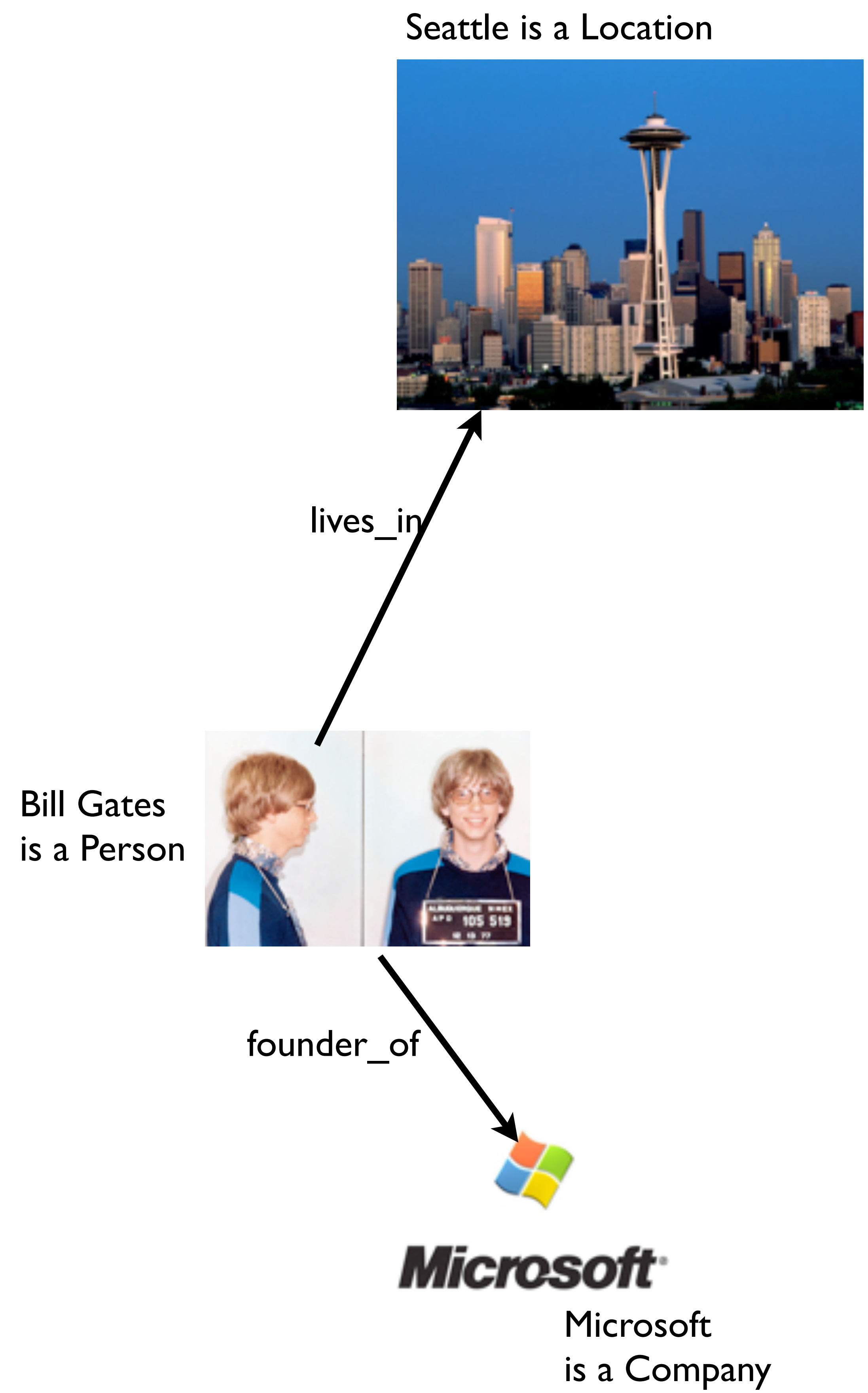
founder_of

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

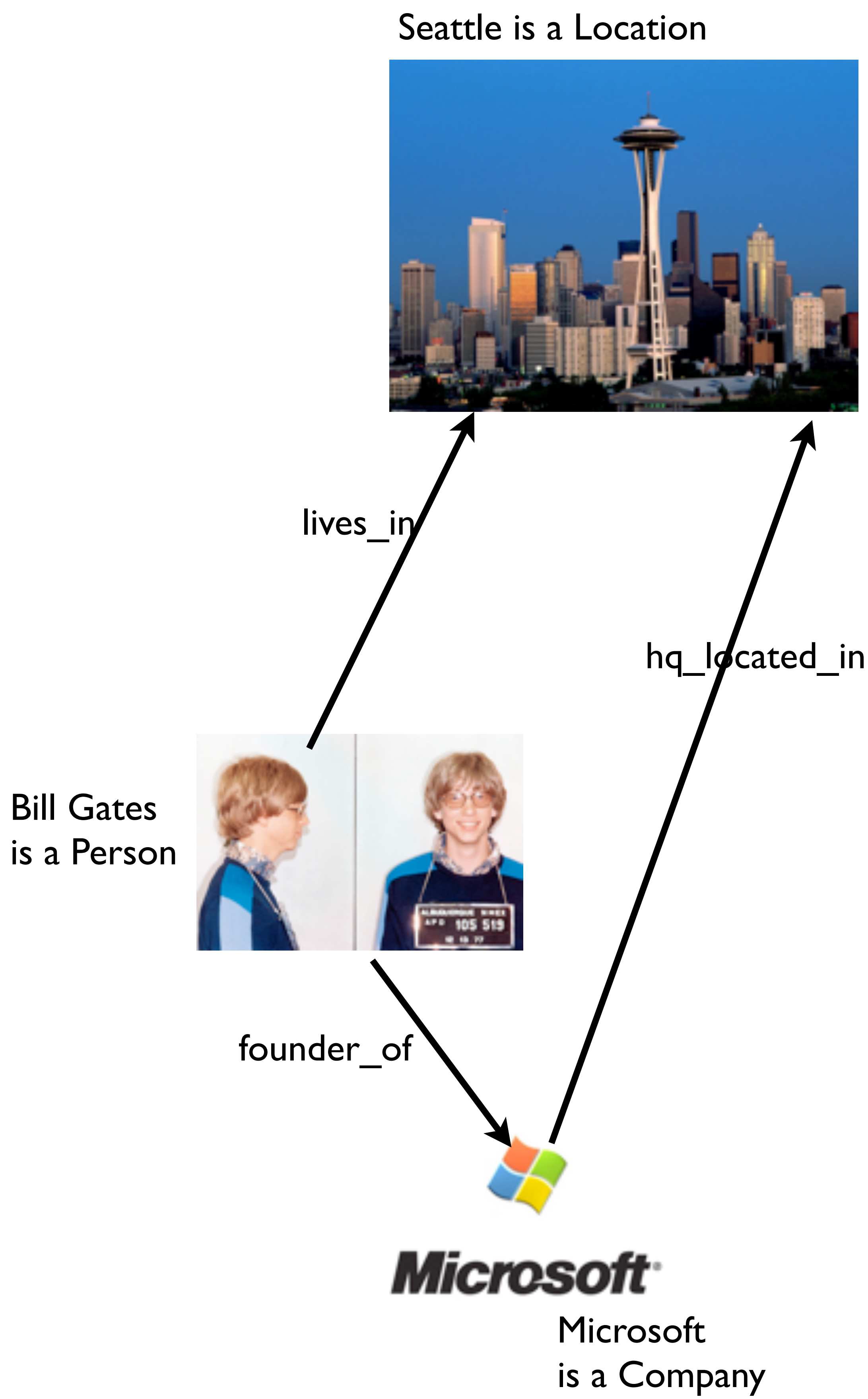
- Early IE-systems were mainly rule-based (manual or learned) and the underlying methodology was specialized for specific applications, cf. MUC systems of the 90th.
- One result of the MUC challenges was a systematic division of labor into IE subtasks
  - Named-Entity Extraction (NER)
  - Relation Entity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

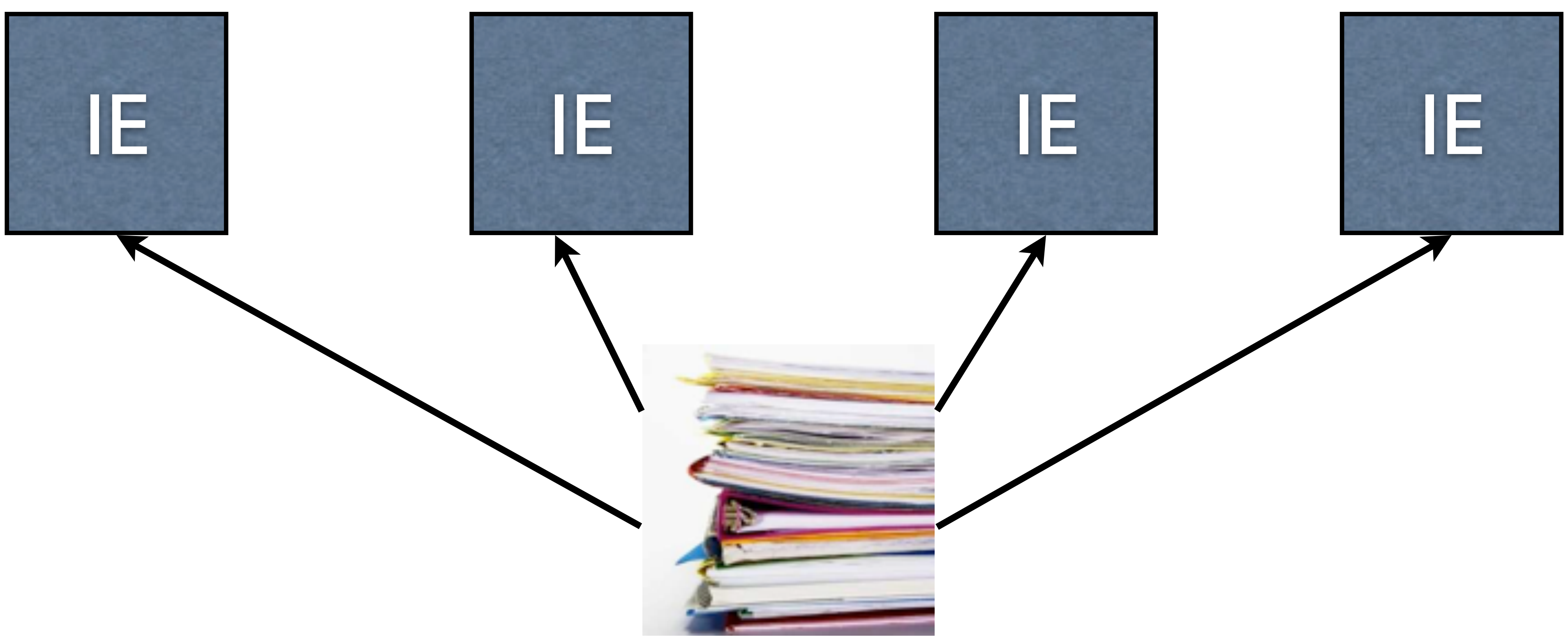
- There exists knowledge-based IE (KIE) and statistical IE (SIE)
- SIE is the State-of-the-Art in research, WIE in industry
- There exists a number of different strategies for the various IE-subtasks
  - from simple gazetteers to complex ontologies
  - from supervised, to minimal supervised to unsupervised 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

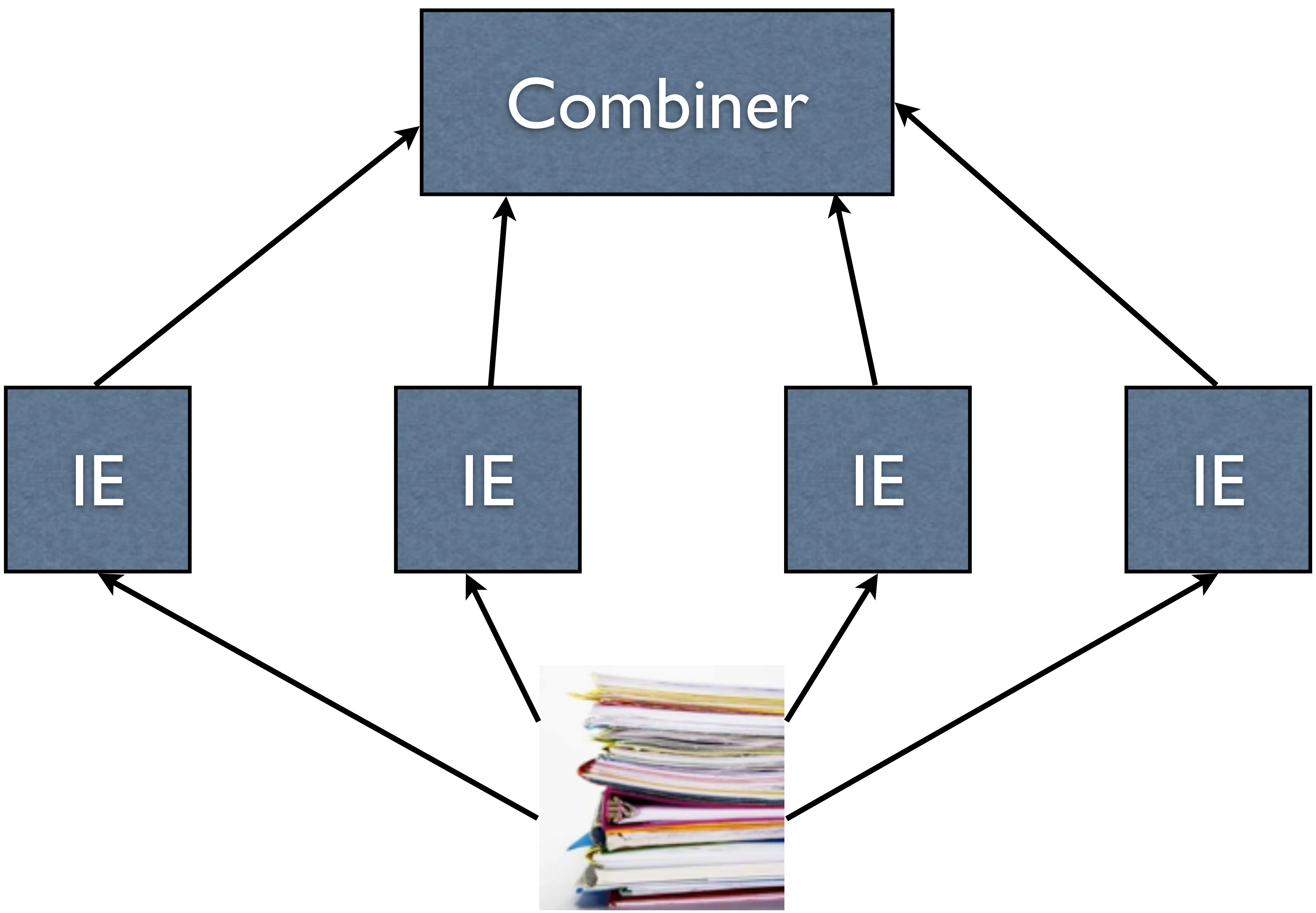
- Methods and strategies for the combination of different IE-components and the analysis of their plausibility.
- What are possible combinations ?

Multi-Strategy

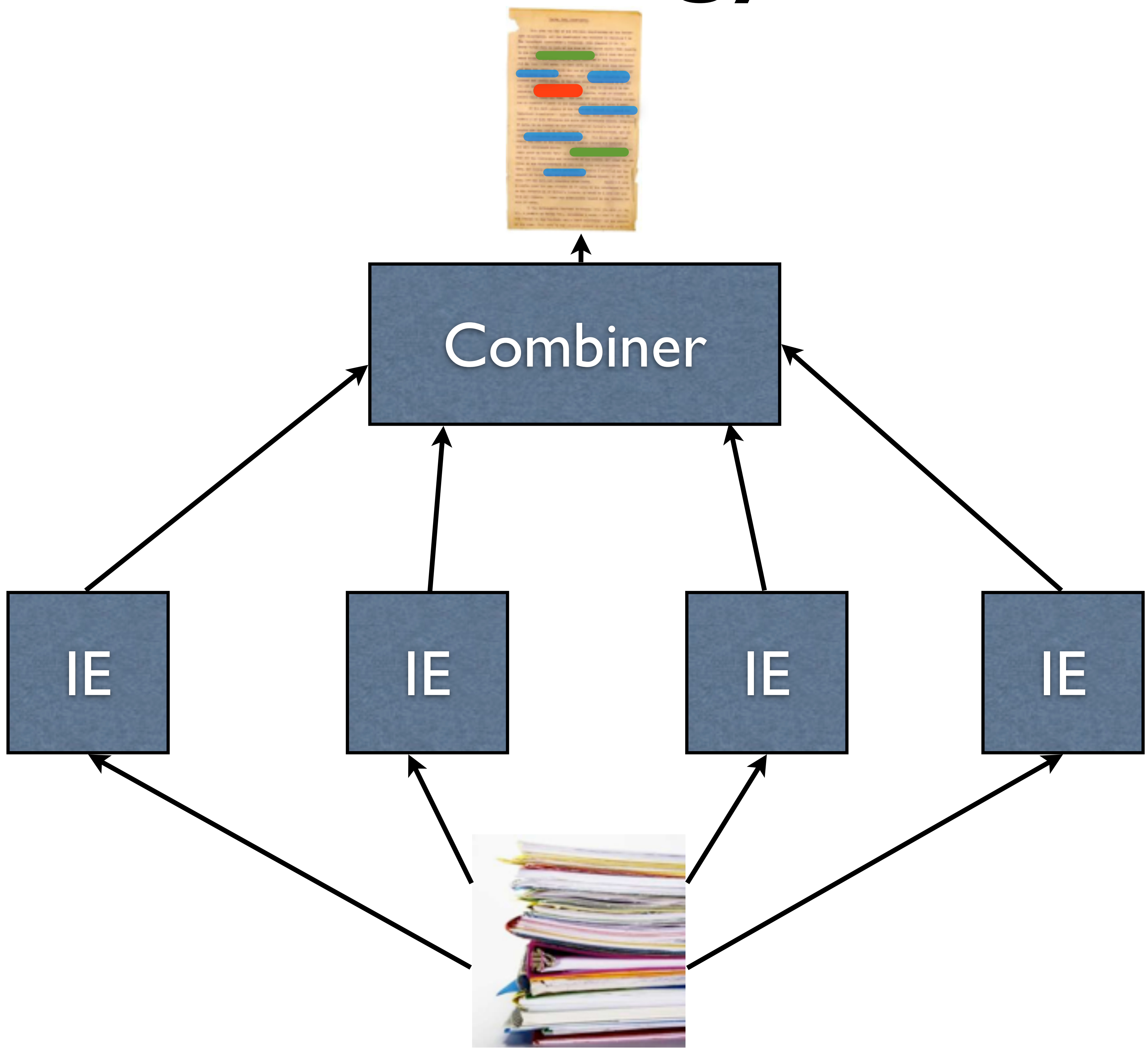
Multi-Strategy



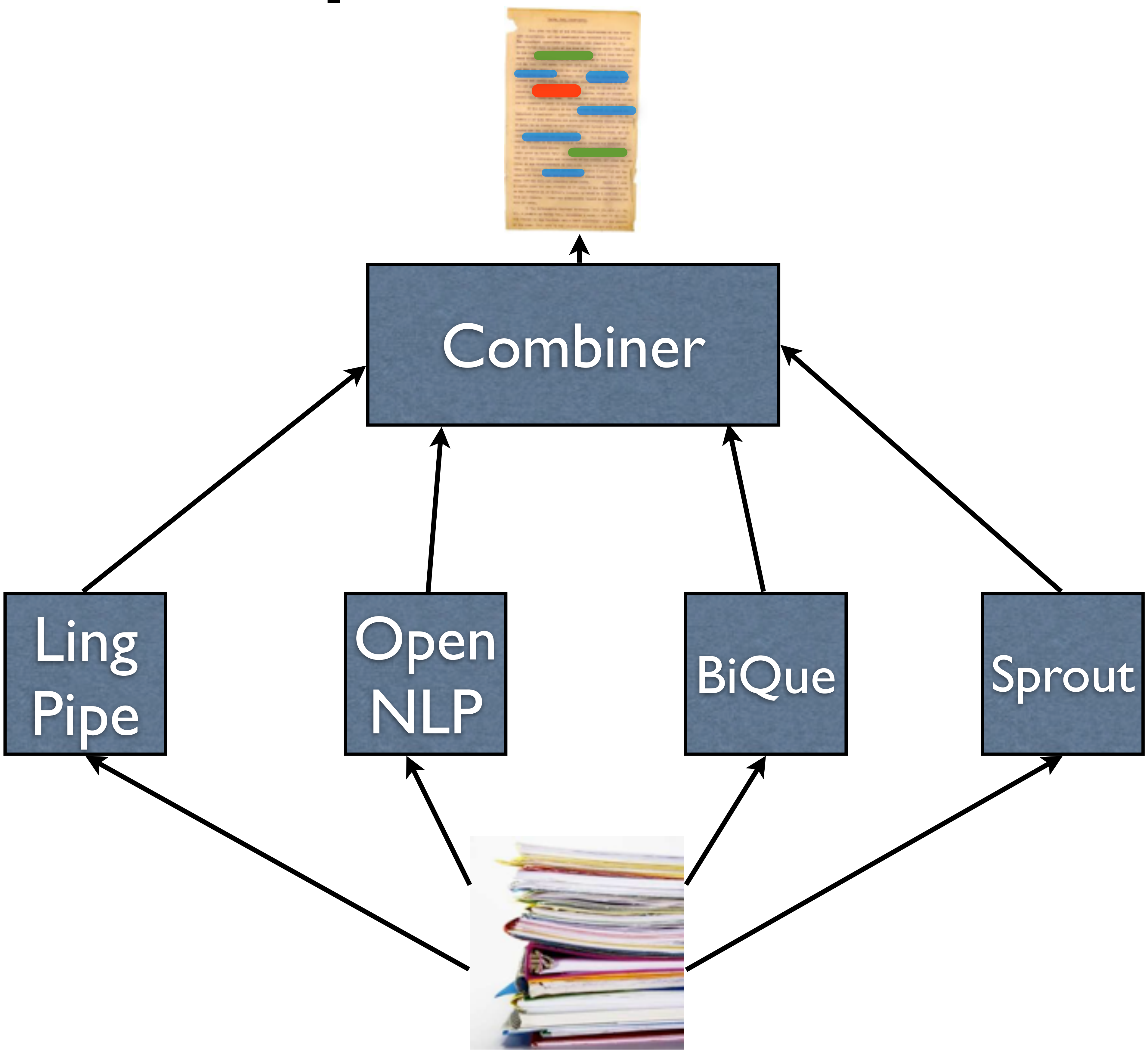
Multi-Strategy



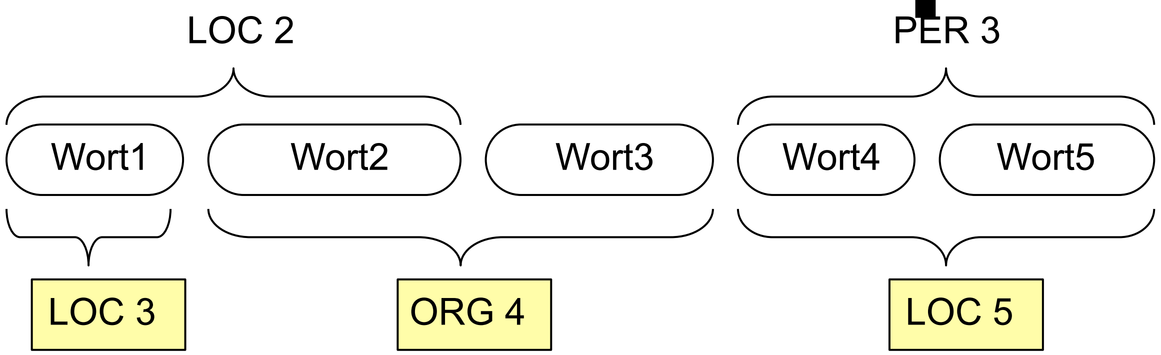
Multi-Strategy



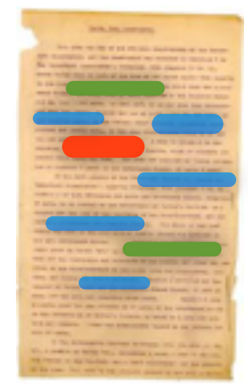
Example: NER

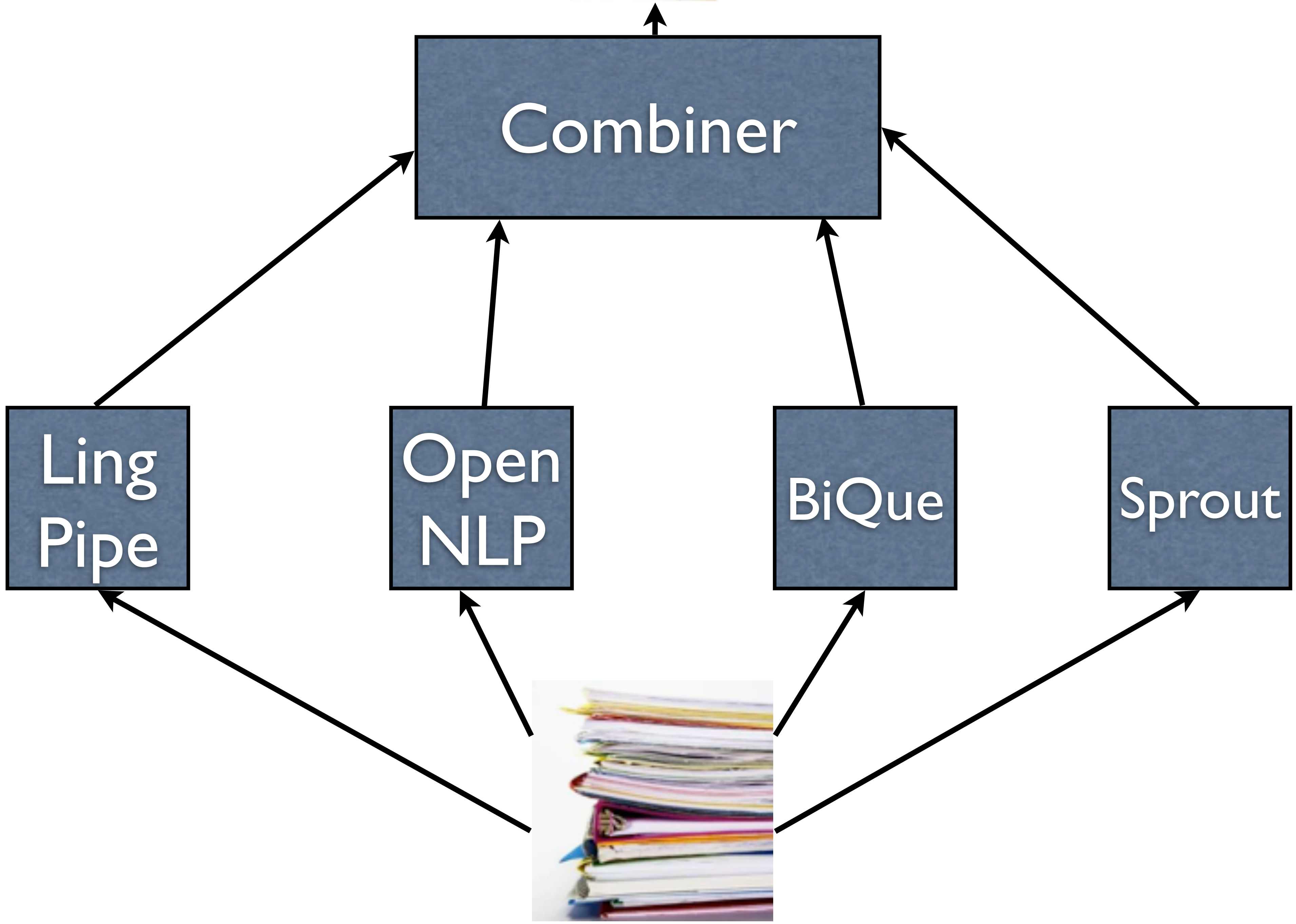


Example: NER

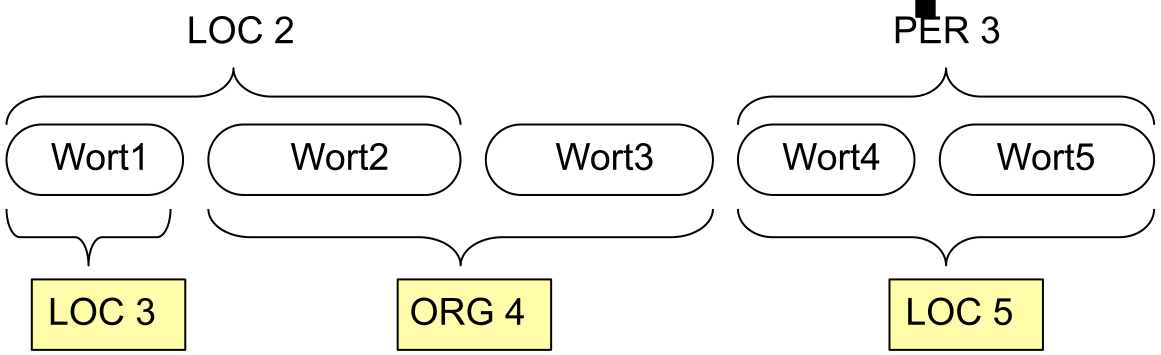


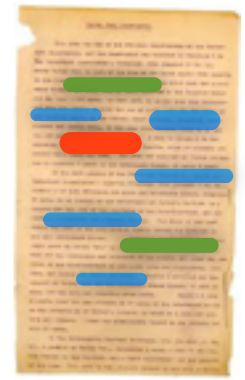
Problem:
- Ambiguities
- Bracketing





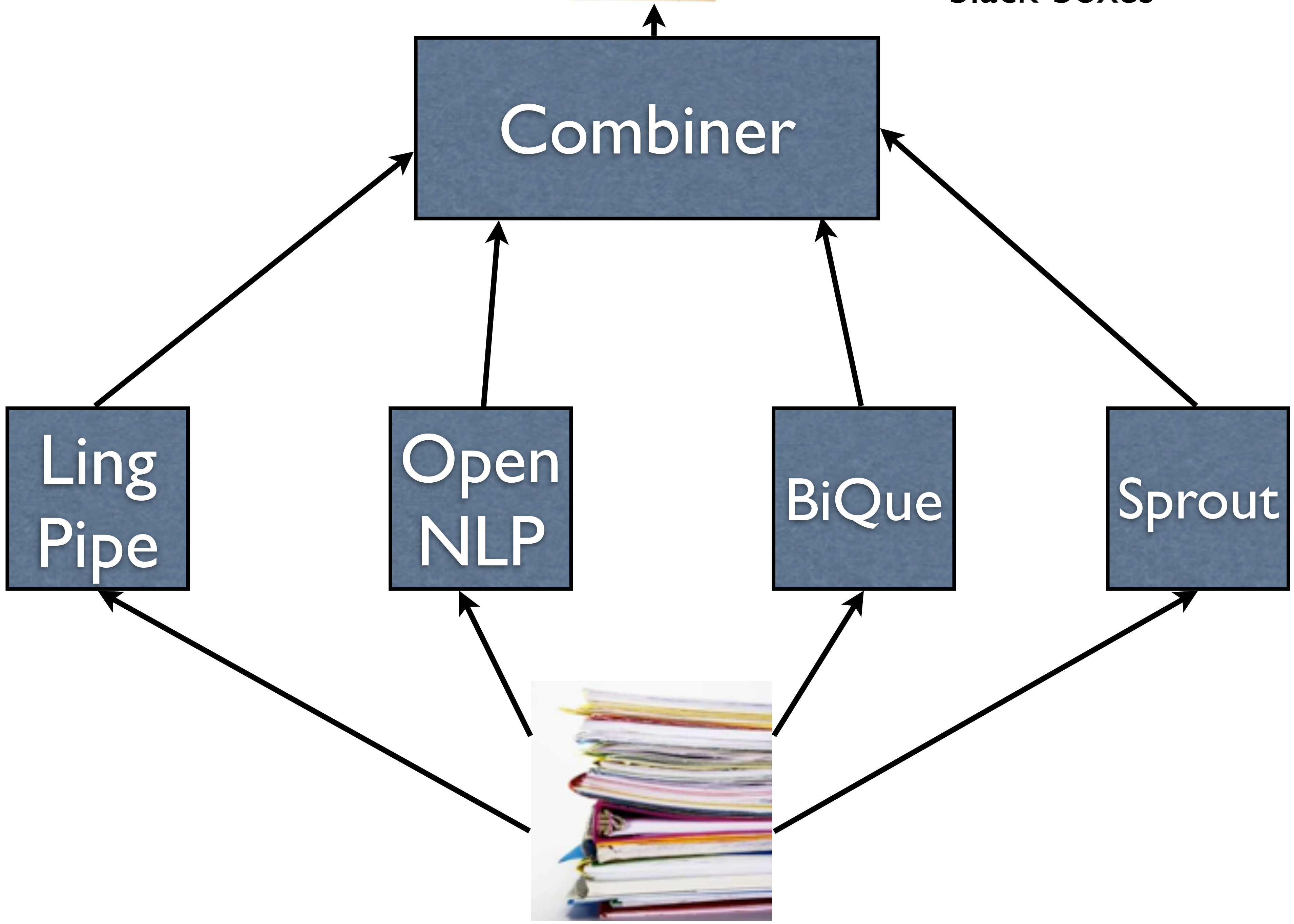
Example: NER



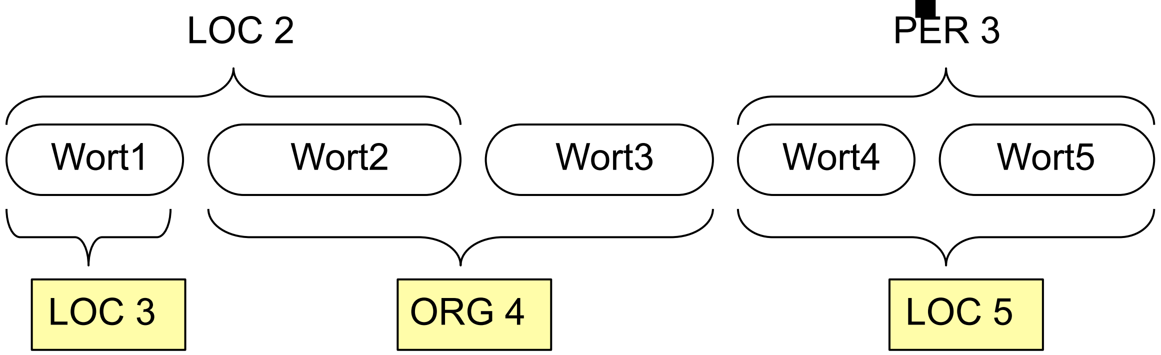


Solutions:
- meta-learning
- consider IE_i as independent black-boxes

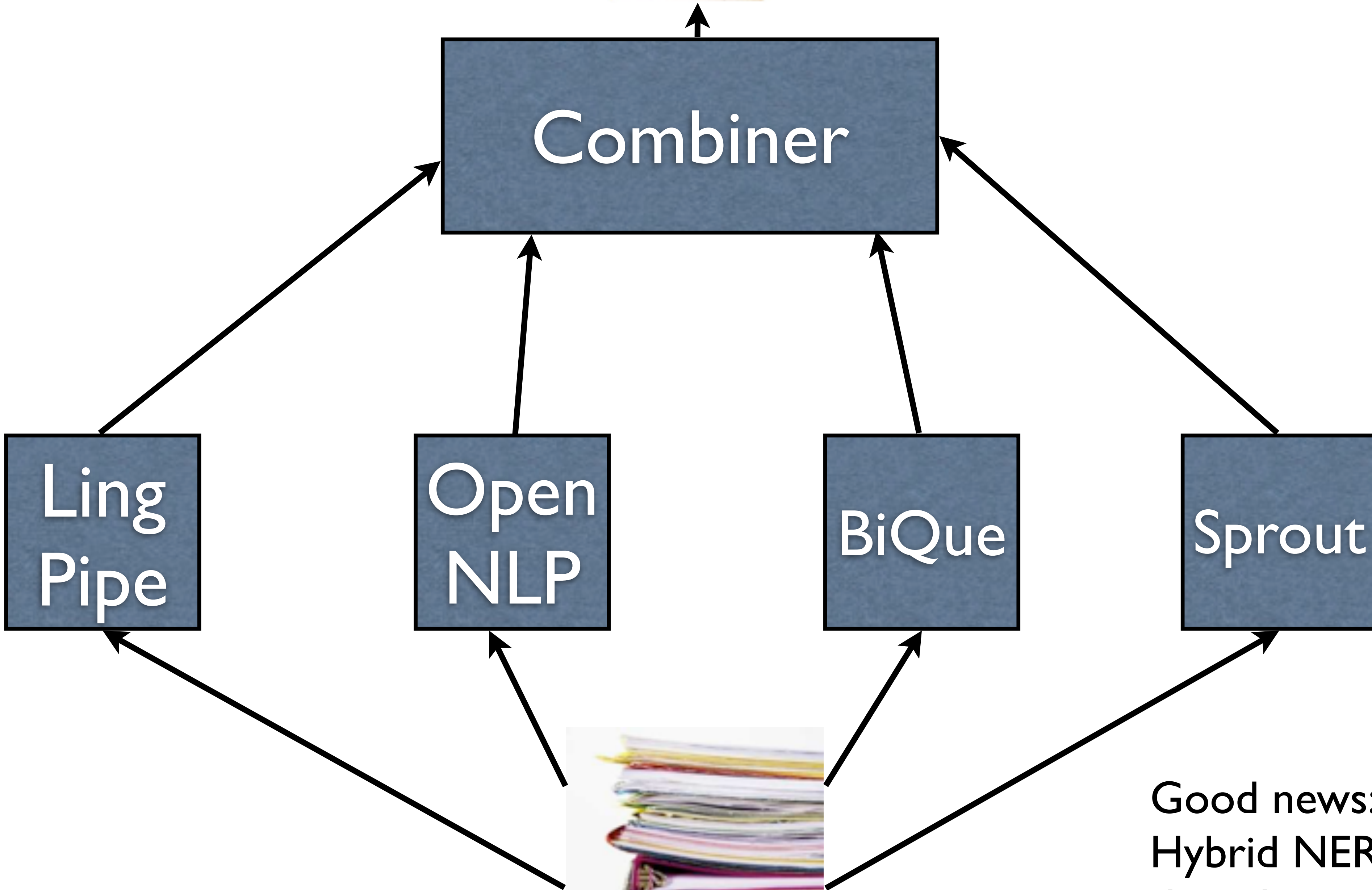
Problem:
- Ambiguities
- Bracketing



Example: NER



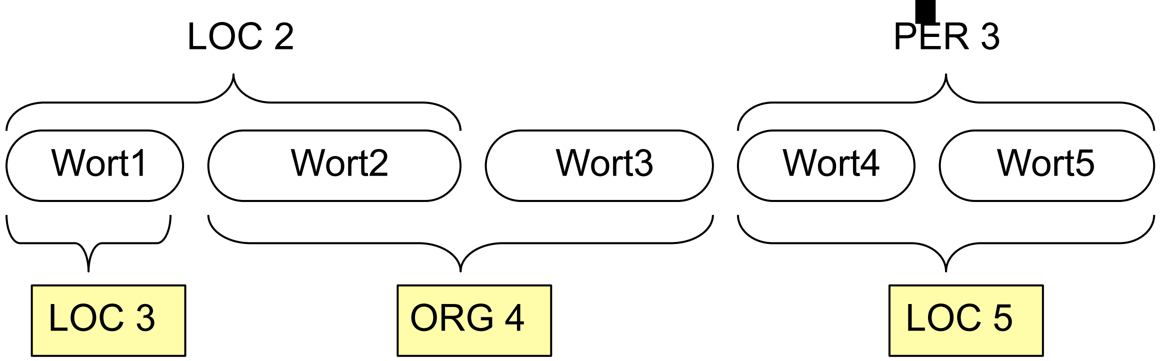
Problem:
- Ambiguities
- Bracketing

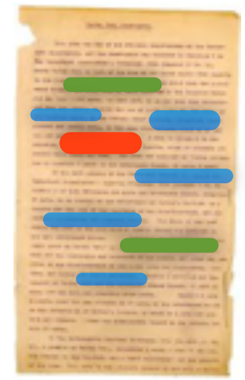


Good news:*
Hybrid NER are better than the single NER wrt. recall and precision.

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

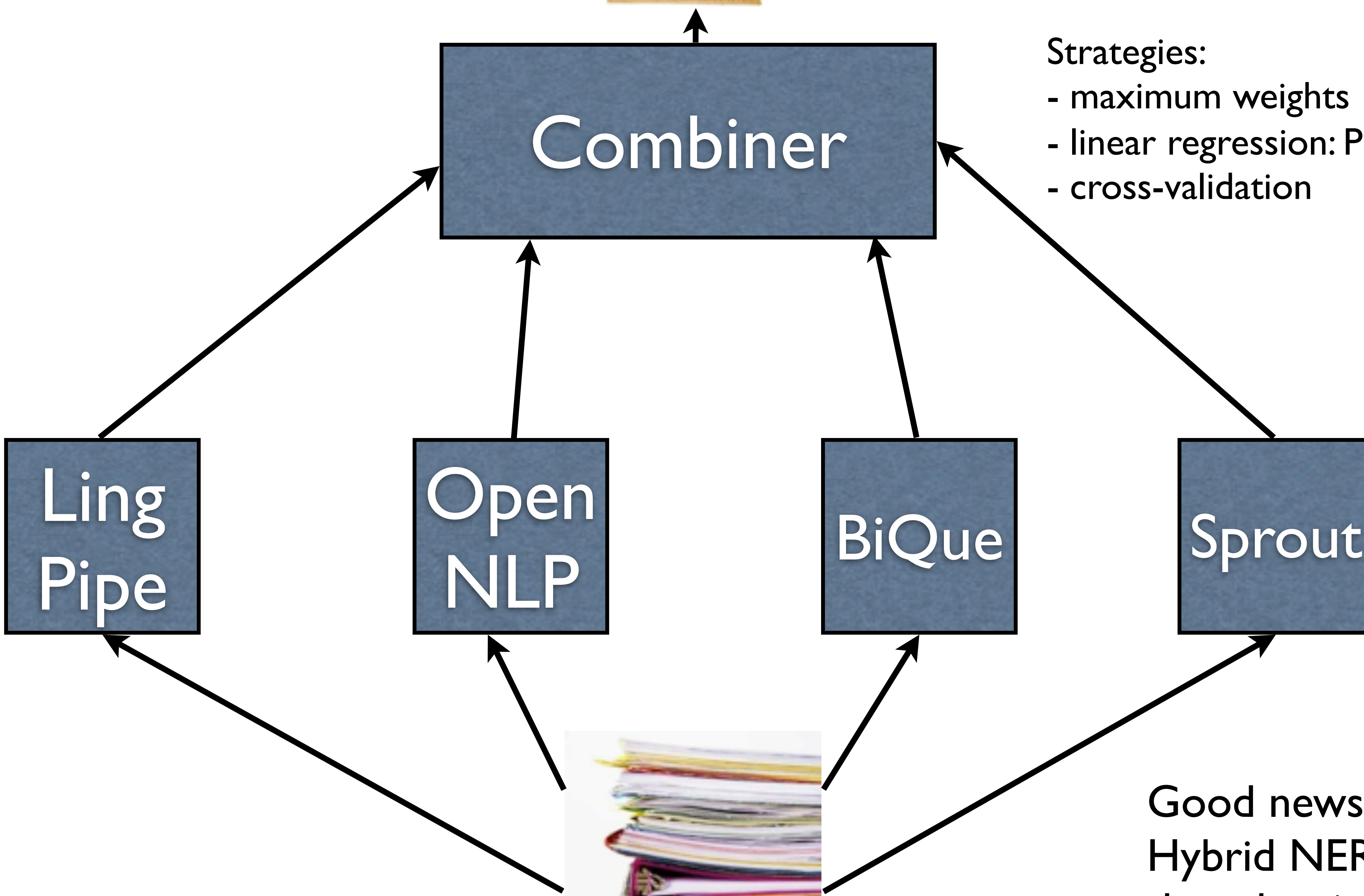




Meta learning
- majority voting
- stacking

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

Problem:
- Ambiguities
- Bracketing

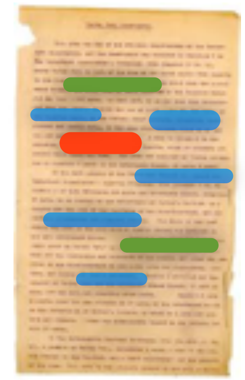


Good news:*
Hybrid NER are better than the single NER wrt. recall and precision.

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: Template Filling

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

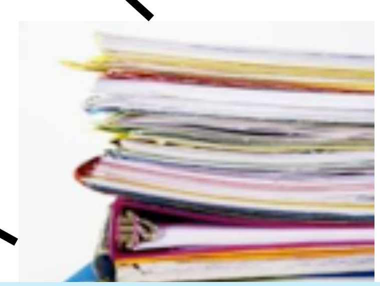


Corpus:
German press releases about turnover (Training: 4850 Tokens, Testing: 1000 Tokens)

Iterative Tag Insertion

MEM - Maximum Entropy Modeling

DOP - Data-Oriented Parsing



Neumann, G. (2006) A Hybrid Machine Learning Approach for Information Extraction from Free Texts. In Spiliopoulou at 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, Heidelber, New-York.

Example: Template Filling

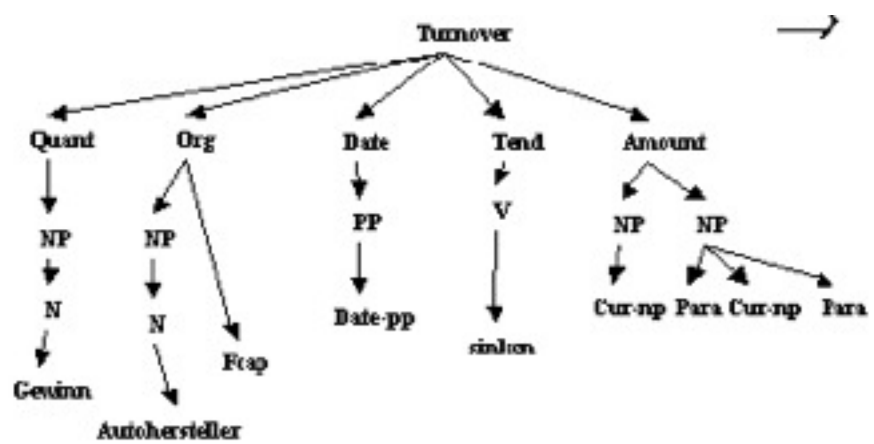
Der Gewinn <Org>der Schweppes Gmbh & Co.</Org> KG betrug <TIMEX>im ersten Quartal 1997</TIMEX> weit ueber 20 Mio. DM.

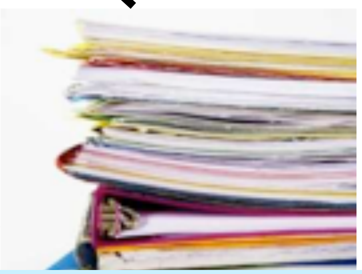
Corpus:
German press releases about turnover (Training: 4850 Tokens, Testing: 1000 Tokens)

Iterative Tag Insertion

MEM - Maximum Entropy Modeling

DOP - Data-Oriented Parsing





Neumann, G. (2006) A Hybrid Machine Learning Approach for Information Extraction from Free Texts. In Spiliopoulou at 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, Heidelber, New-York.

Example: Template Filling

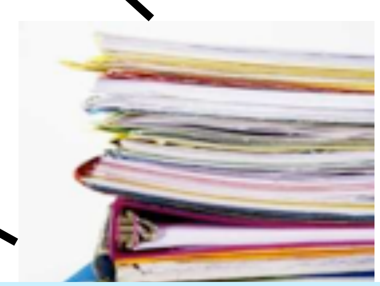
Der Gewinn <Org>der Schweppes Gmbh & Co.</Org> KG betrug <TIMEX>im ersten Quartal 1997</TIMEX> weit ueber 20 Mio. DM.

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

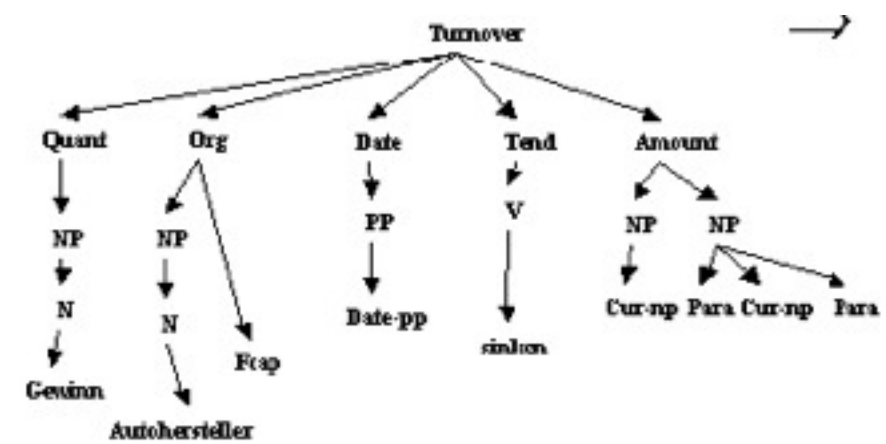
Iterative Tag Insertion

MEM - Maximum Entropy Modeling

DOP - Data-Oriented Parsing



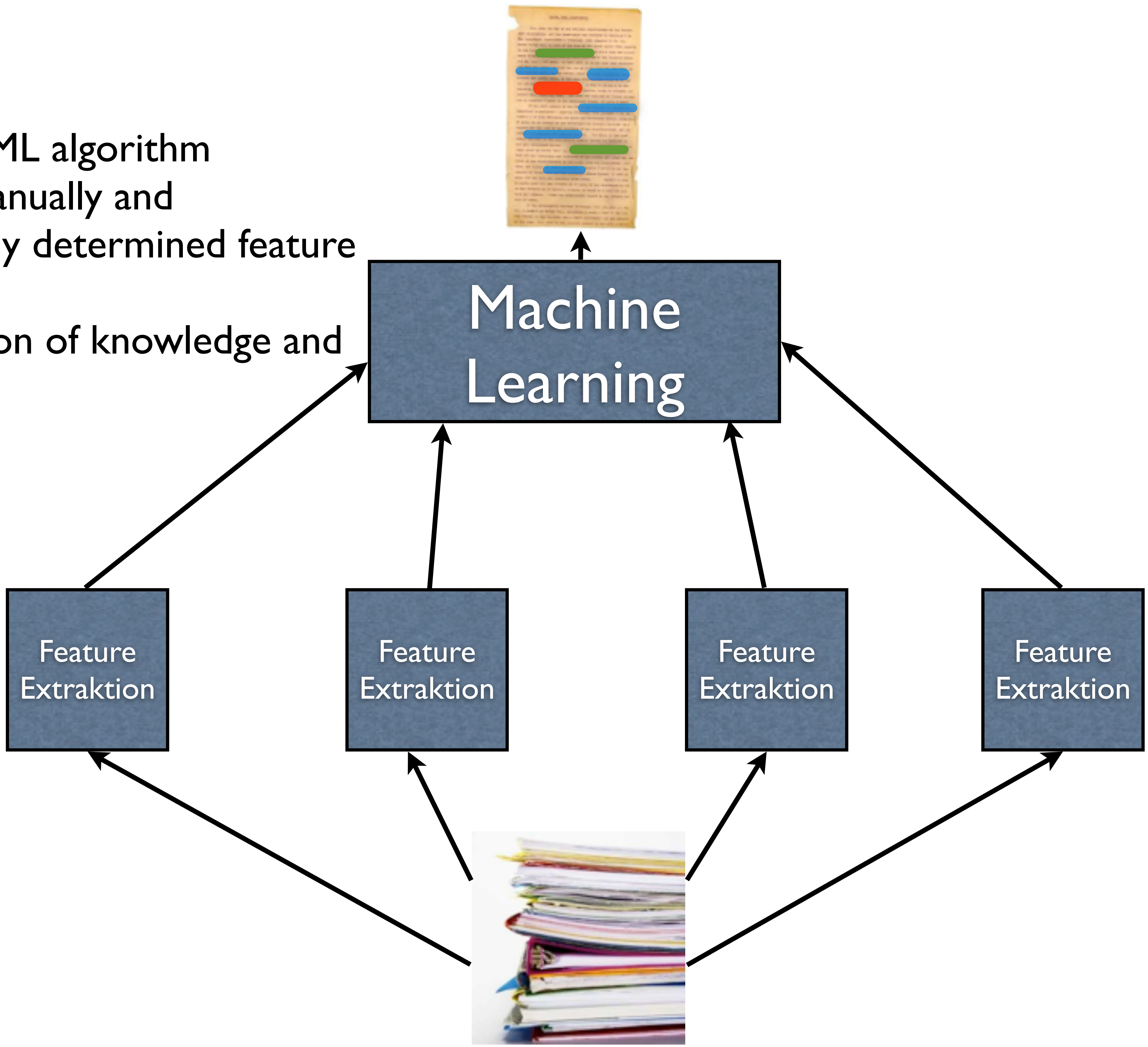
Result:
- only MEM: 79.3 %
- only DOP: 51.9 %
- both: 85.2 %



Neumann, G. (2006) A Hybrid Machine Learning Approach for Information Extraction from Free Texts. In Spiliopoulou at 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, Heidelber, New-York.

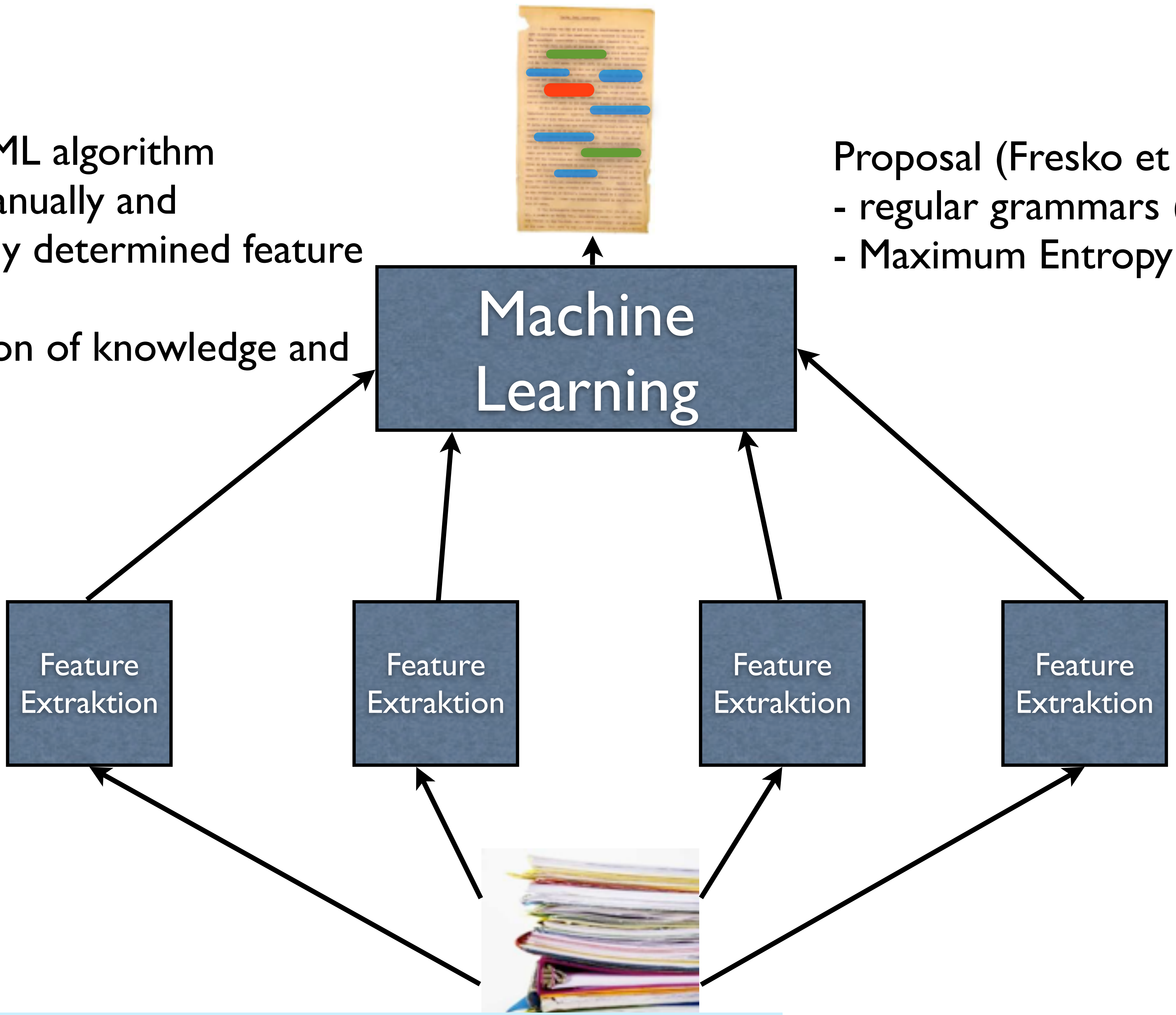
Feature based Strategies

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



Feature based Strategies

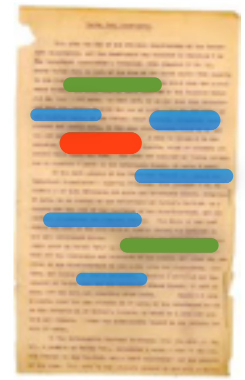
Idea:
- choose a ML algorithm
- choose manually and automatically determined feature templates
- combination of knowledge and statistics



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

Machine Learning

Feature Extraktion

Feature Extraktion

Feature Extraktion

Feature Extraktion

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**

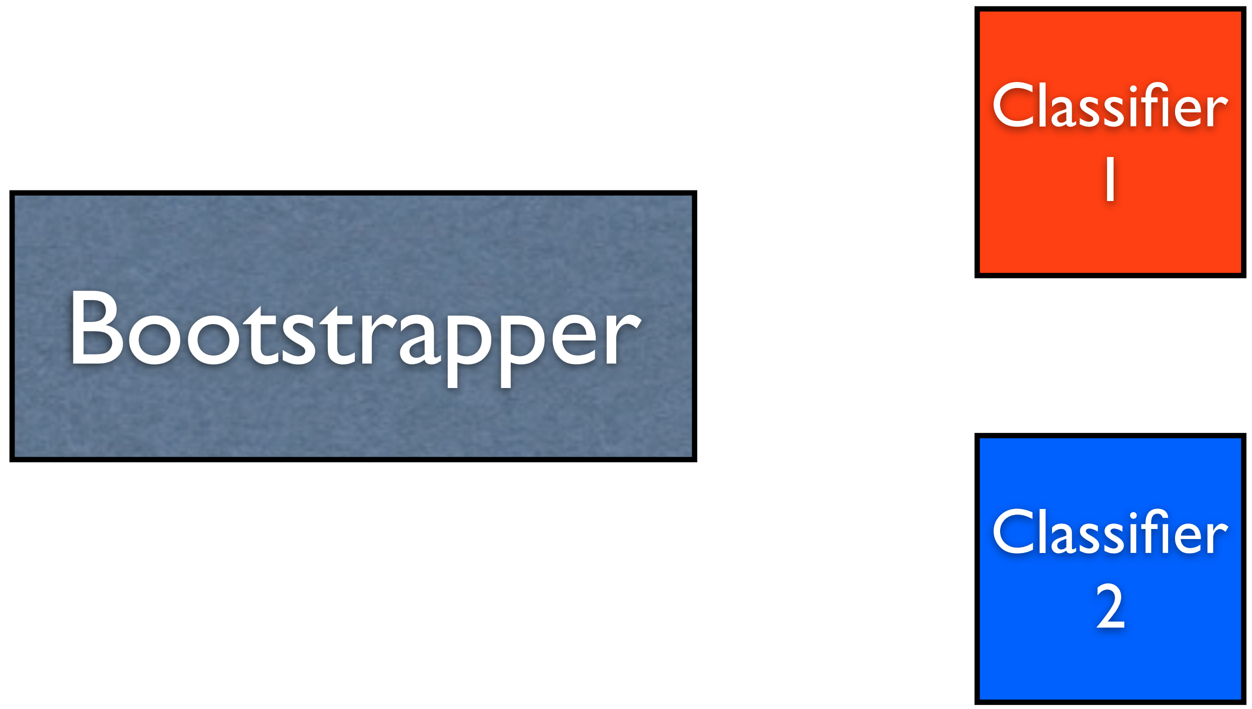
Table 1. MUC7: After training with 100 documents.

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

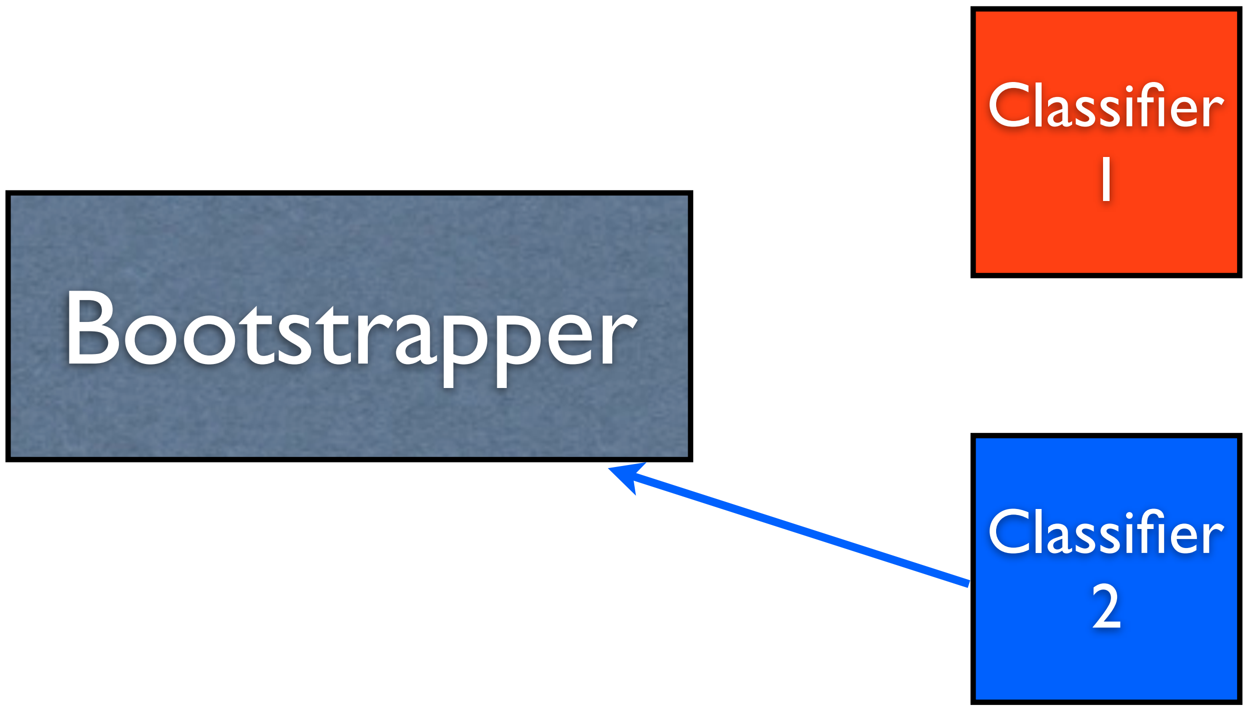
Co-Training & Bootstrapping

Bootstrapper

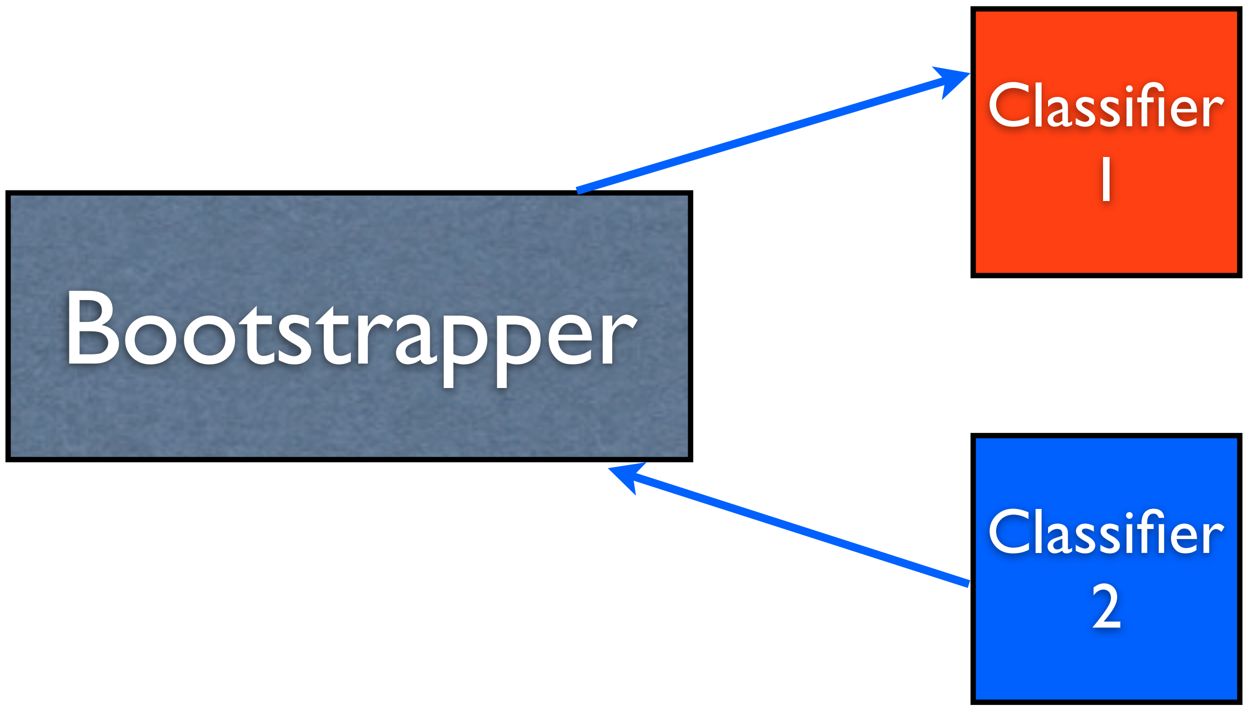
Co-Training & Bootstrapping



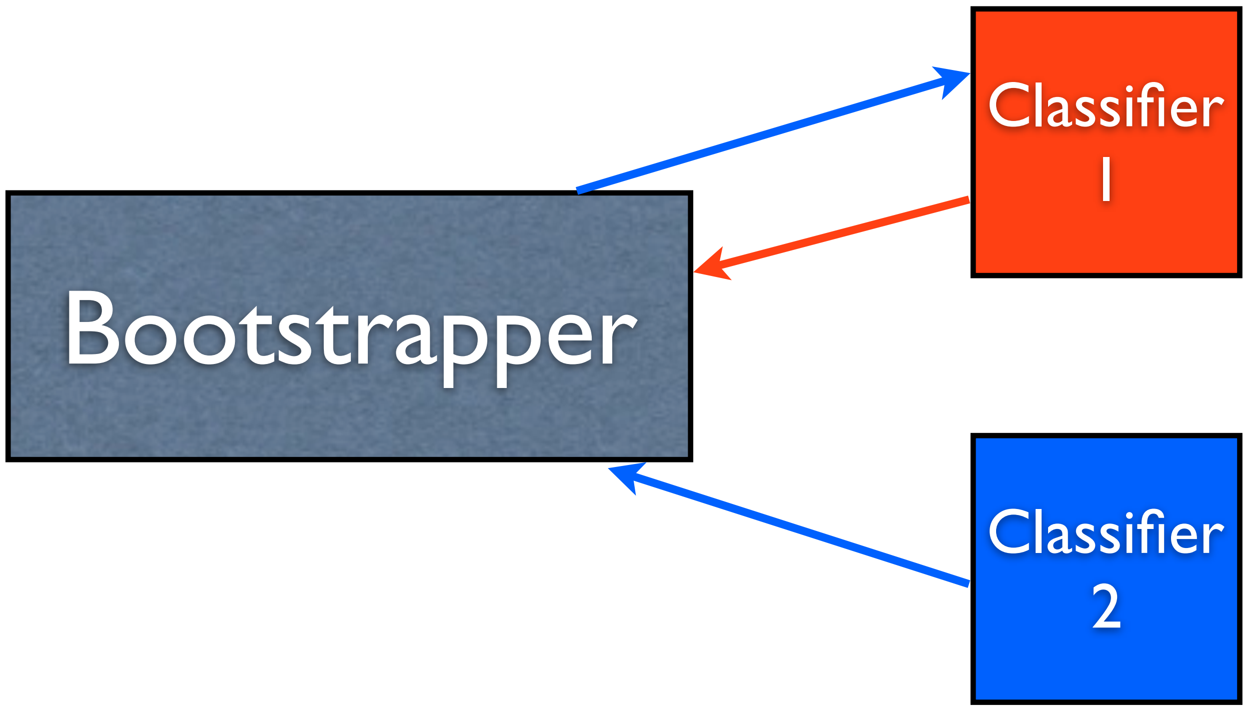
Co-Training & Bootstrapping



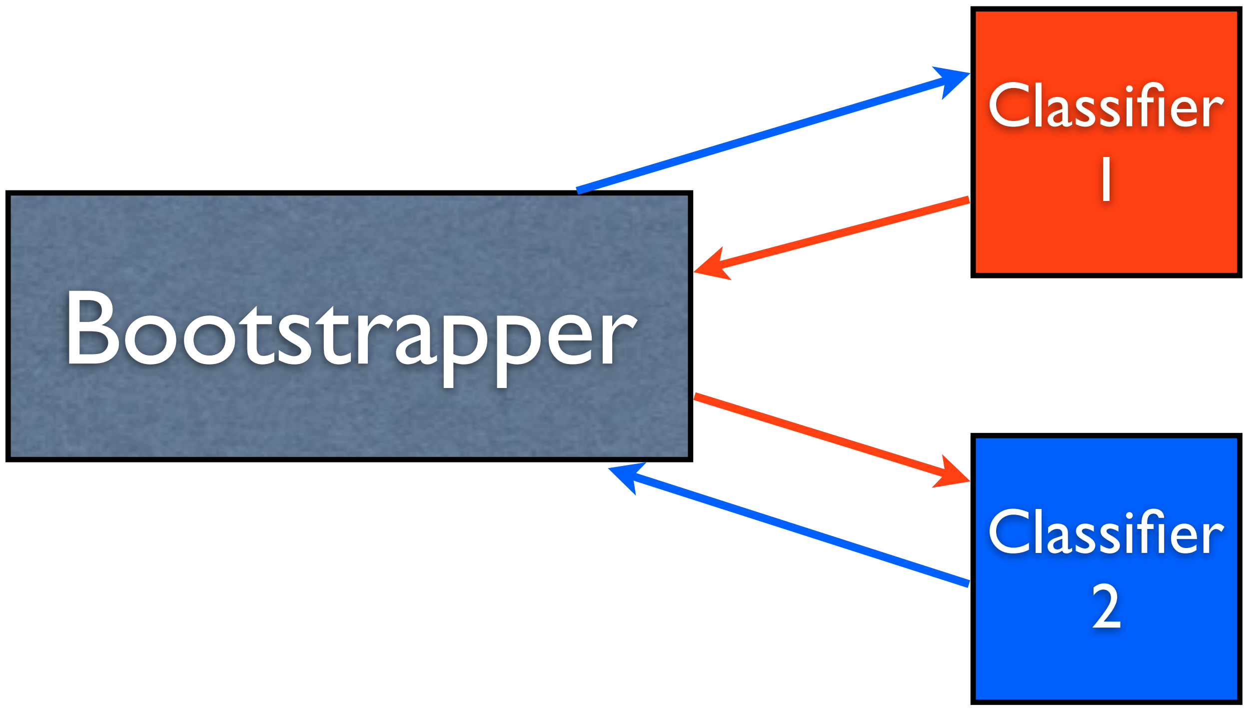
Co-Training & Bootstrapping



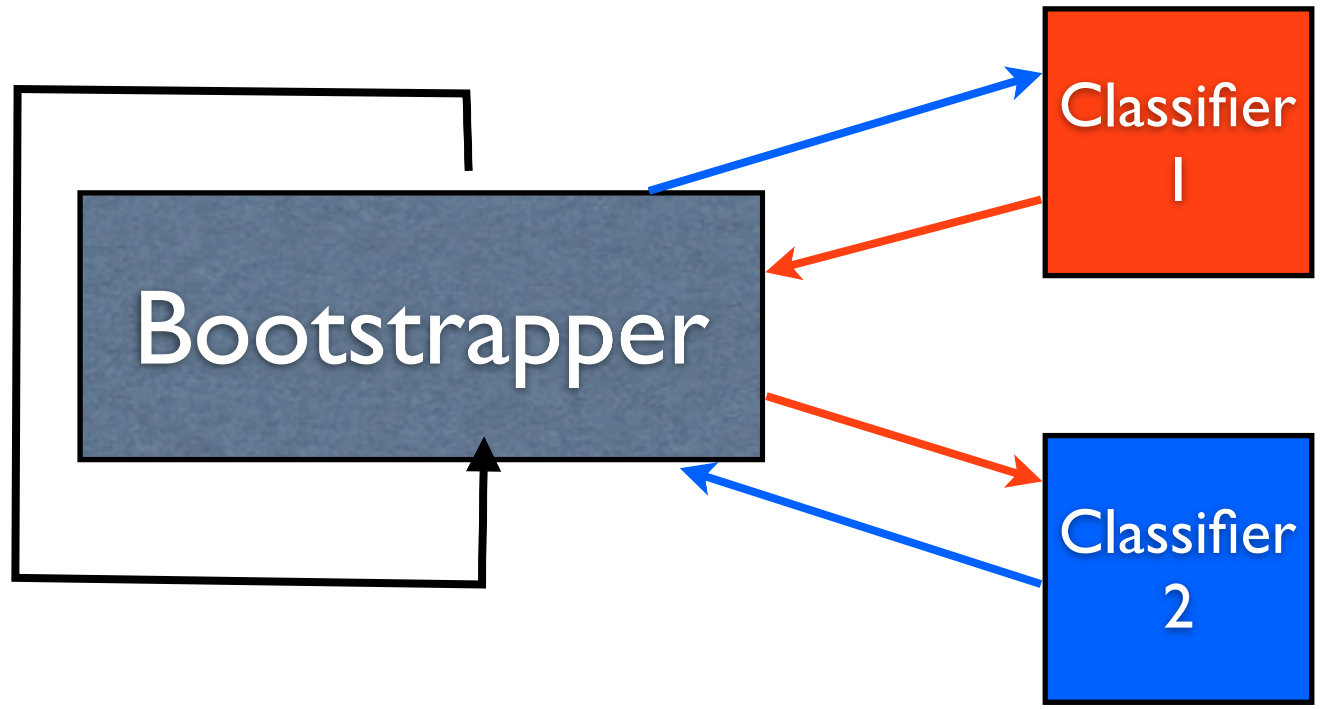
Co-Training & Bootstrapping



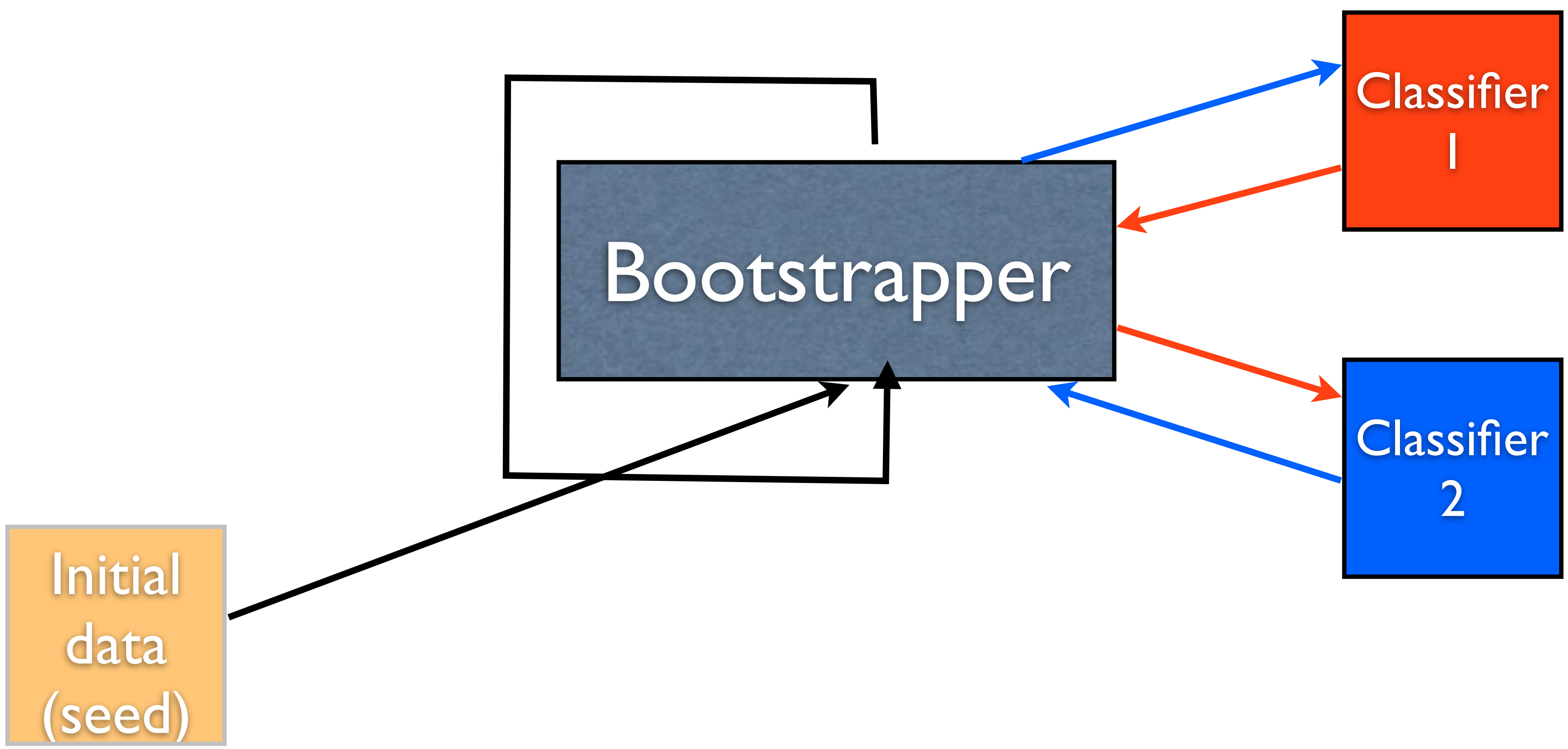
Co-Training & Bootstrapping



Co-Training & Bootstrapping

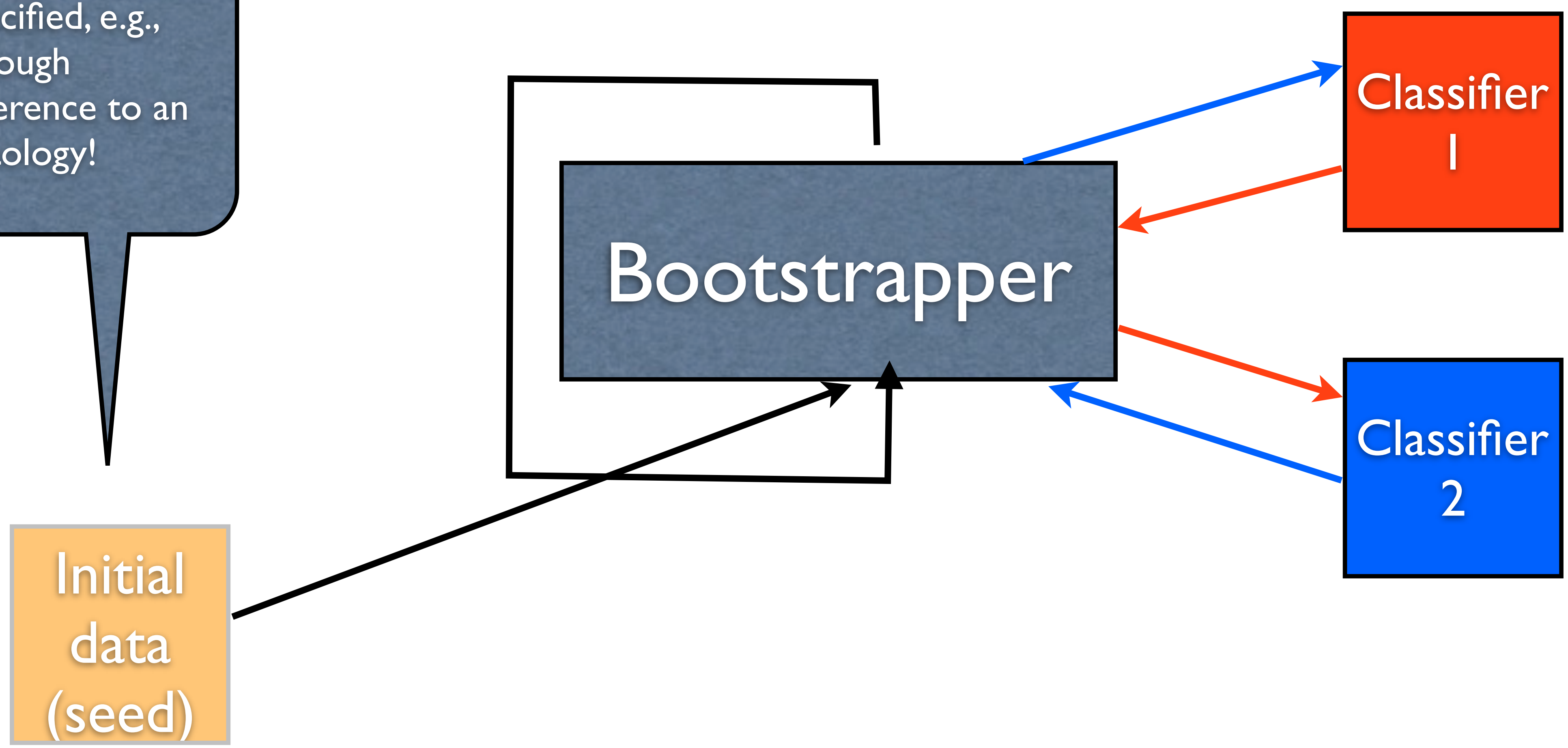


Co-Training & Bootstrapping



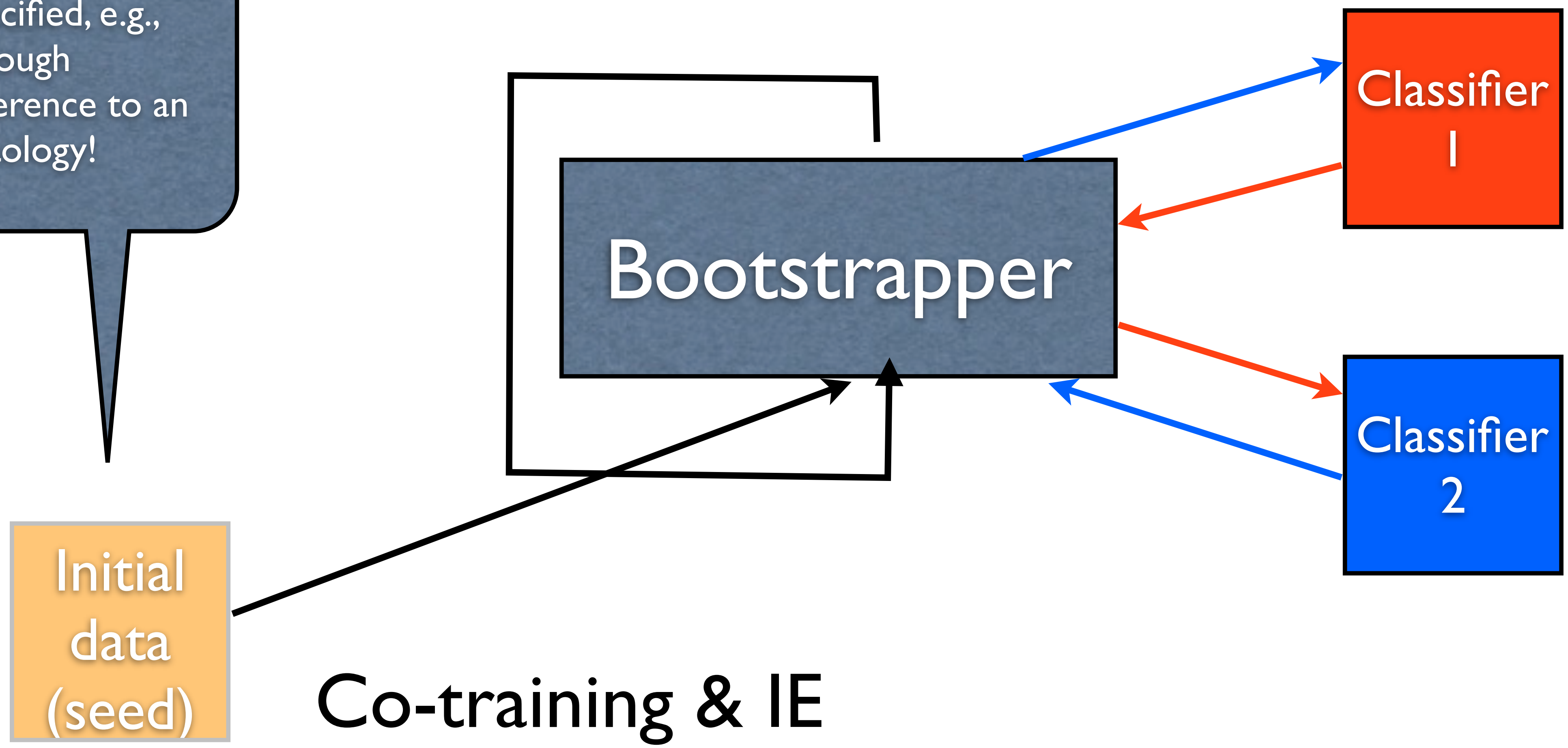
Co-Training & Bootstrapping

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

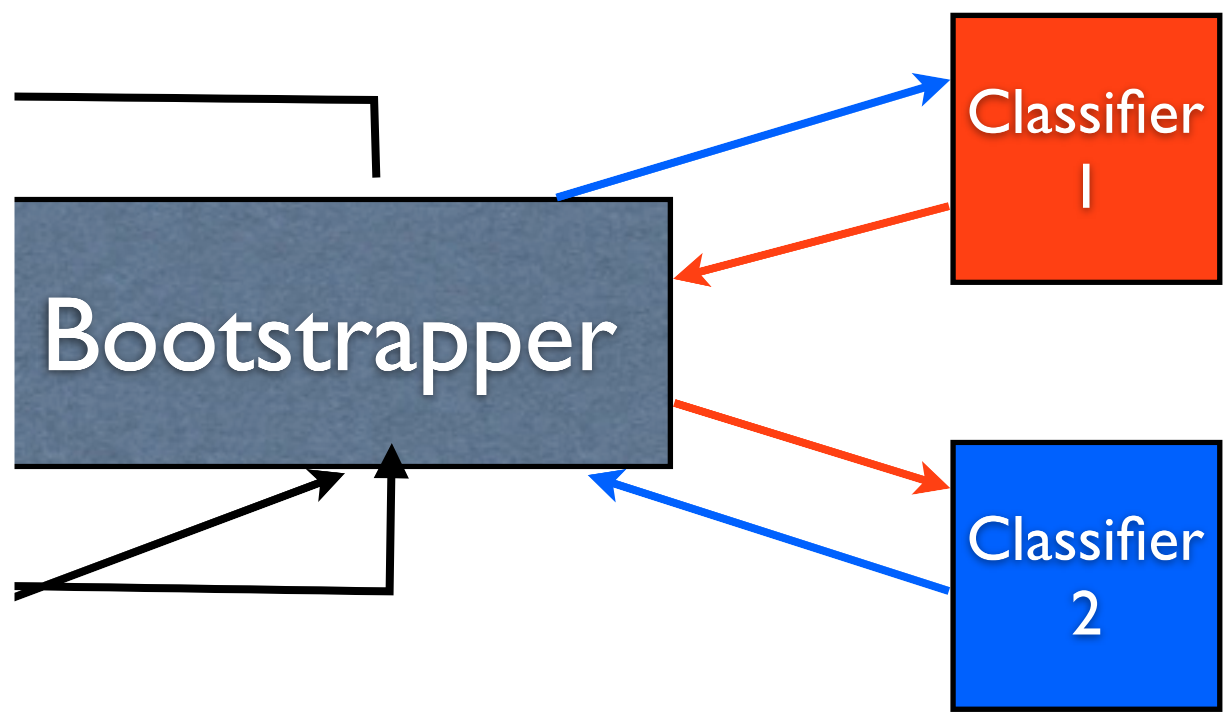
- NER, cf Singer & Collins, 1999
  Interaction of spelling and context features
- REE, cf. Surdeanu et al. 2006
  Interaction of text classifier and pattern acquisition

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:?)

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.

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

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.

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

Was ist XYZ ?

XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a

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.

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!

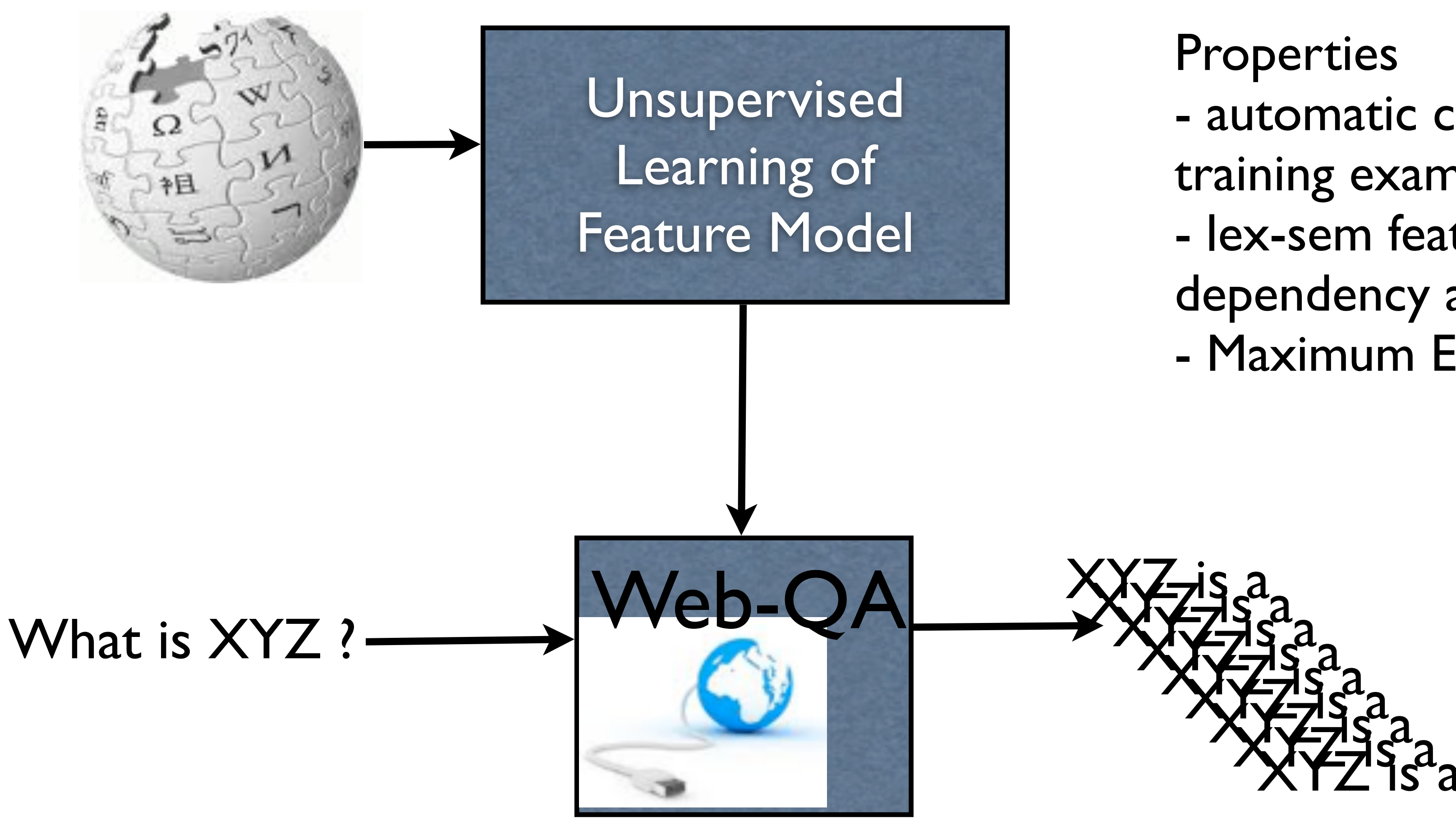
Wikipedia as Blueprint!

- 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!

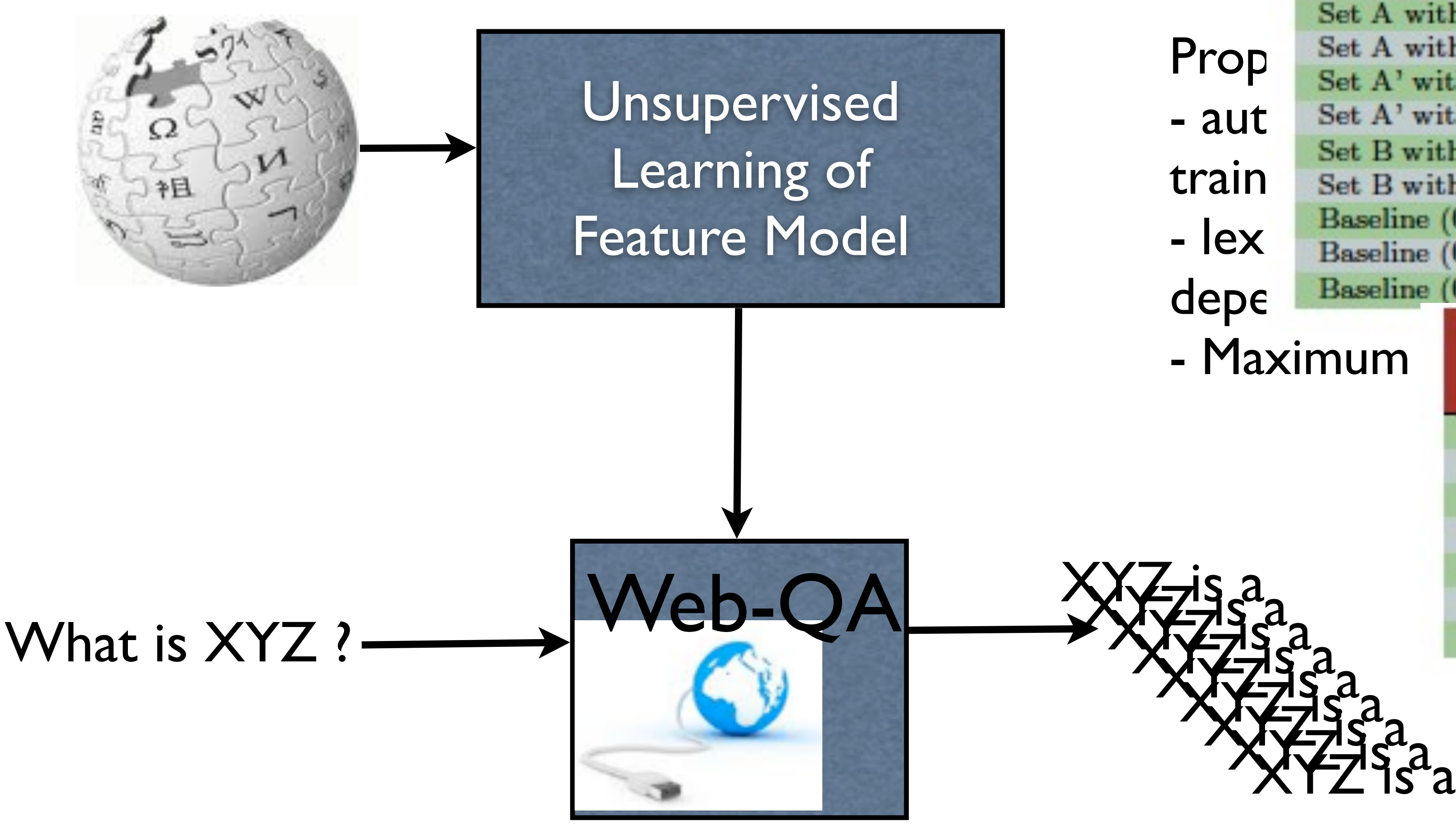


Properties
- automatic computation of POS und NEG training examples
- lex-sem feature-templates via dependency analysis
- Maximum Entropy Modeling

Solution:
Rank answer candidates 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' Accuracy	Applied to Set A 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!

Unsupervised Learning of Feature Model

What is XYZ ?

Web-QA

XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a
XYZ is a

Prop
- aut
train
- lex
depe
- Maximum

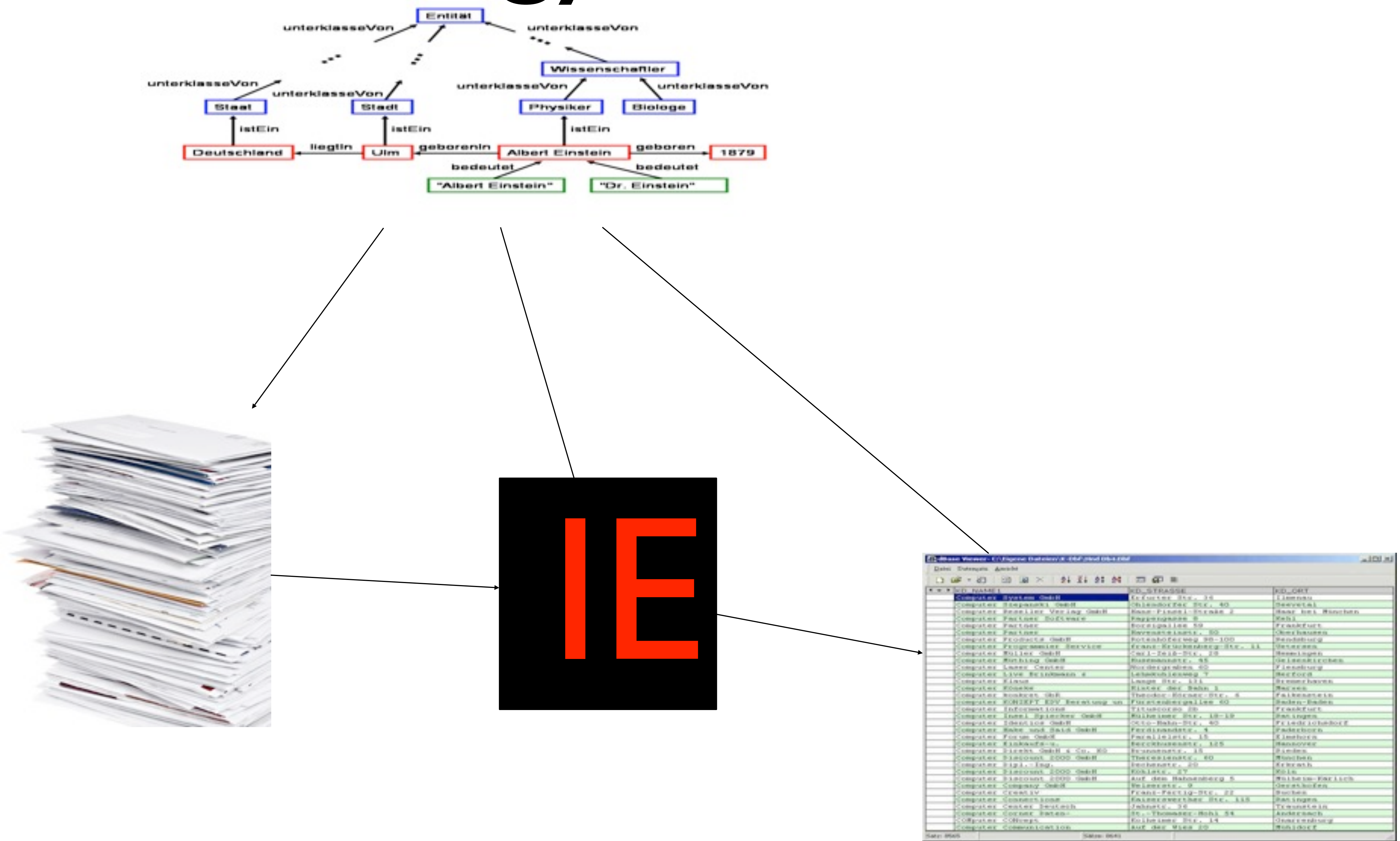
Best Attributes Found for	Applied to Set A' Accuracy	Applied to Set A Accuracy
Set A without NLP	81.16%	
Set A with NLP	**85.94%**	
Set A' without NLP		78.28%
Set A' with NLP		**83.04%**
Set B without NLP	76.86%	75.50%
Set B with NLP	63.19%	74.61%
Baseline (0.3)	54.13%	44.17%
Baseline (0.2)	54.63%	44.23%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

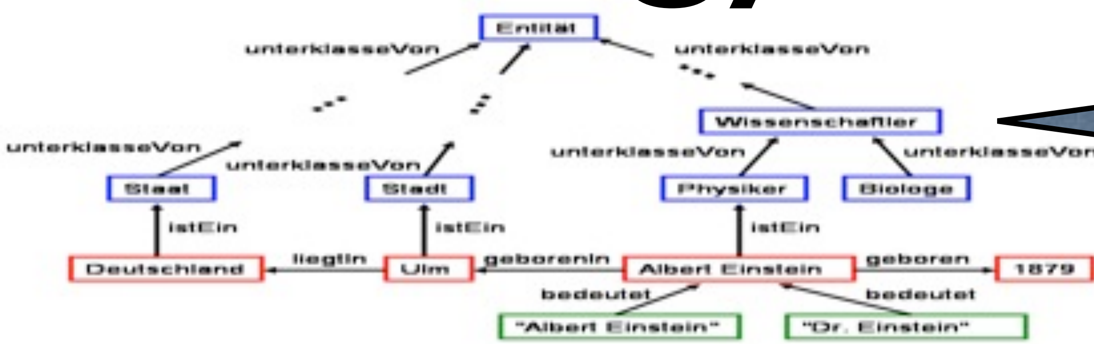
according to similarity of Wikipedia?

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

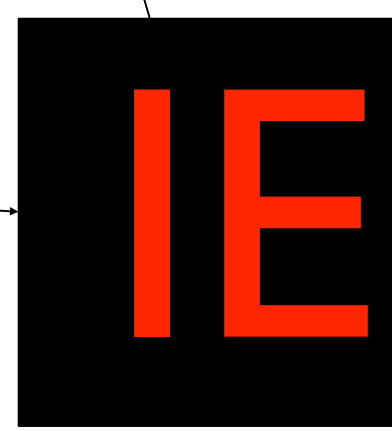
Ontology based IE

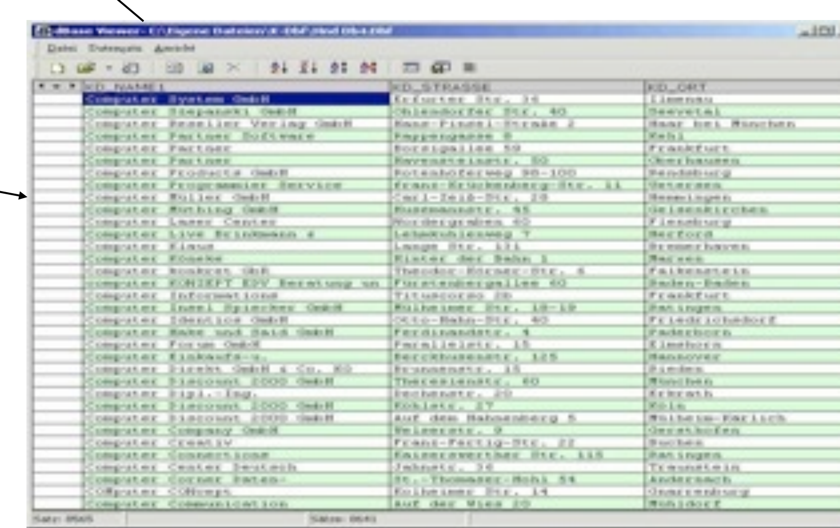


Ontology based IE

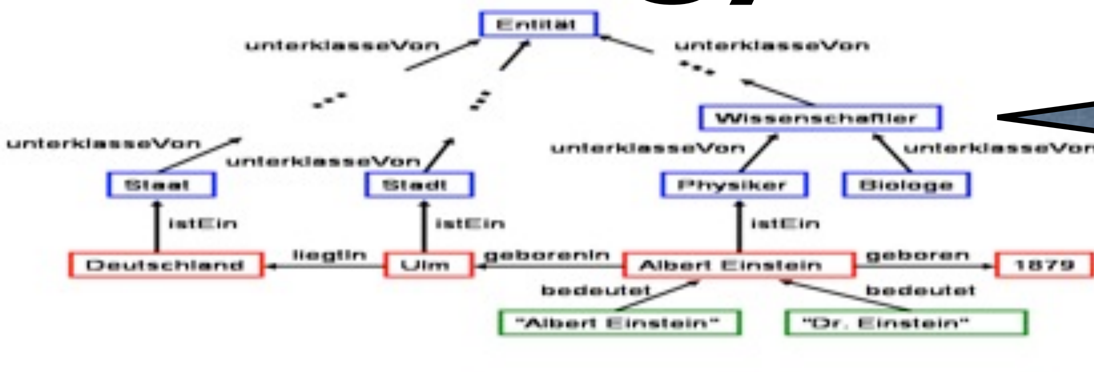


The ontology defines the type of the information, which has to be extracted from texts: e.g., types of or institutions and their inter relationship. It defines the structure of the data base, which has to be extracted automatically with the help of OBIE.



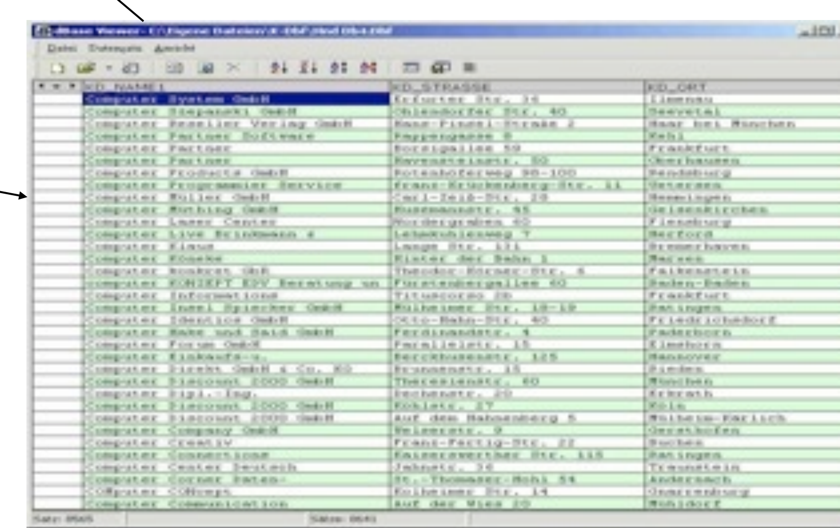


Ontology based IE



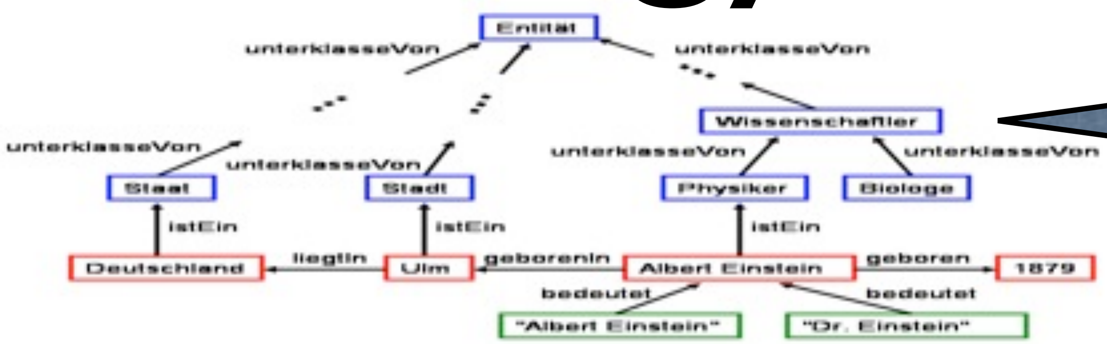
The ontology defines the type of the information, which has to be extracted from texts: e.g., types of or institutions and their inter relationship. It defines the structure of the data base, which has to be extracted automatically with the help of OBIE.

IE



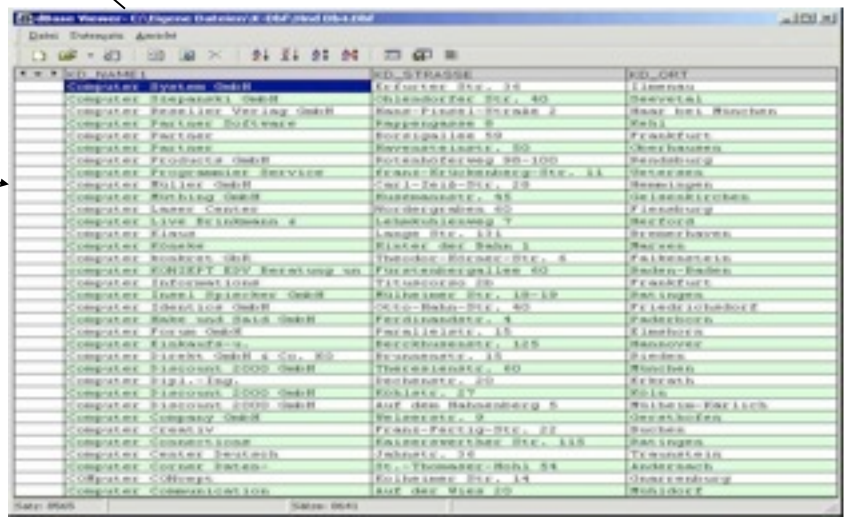
Ontology population
ontology learning

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.

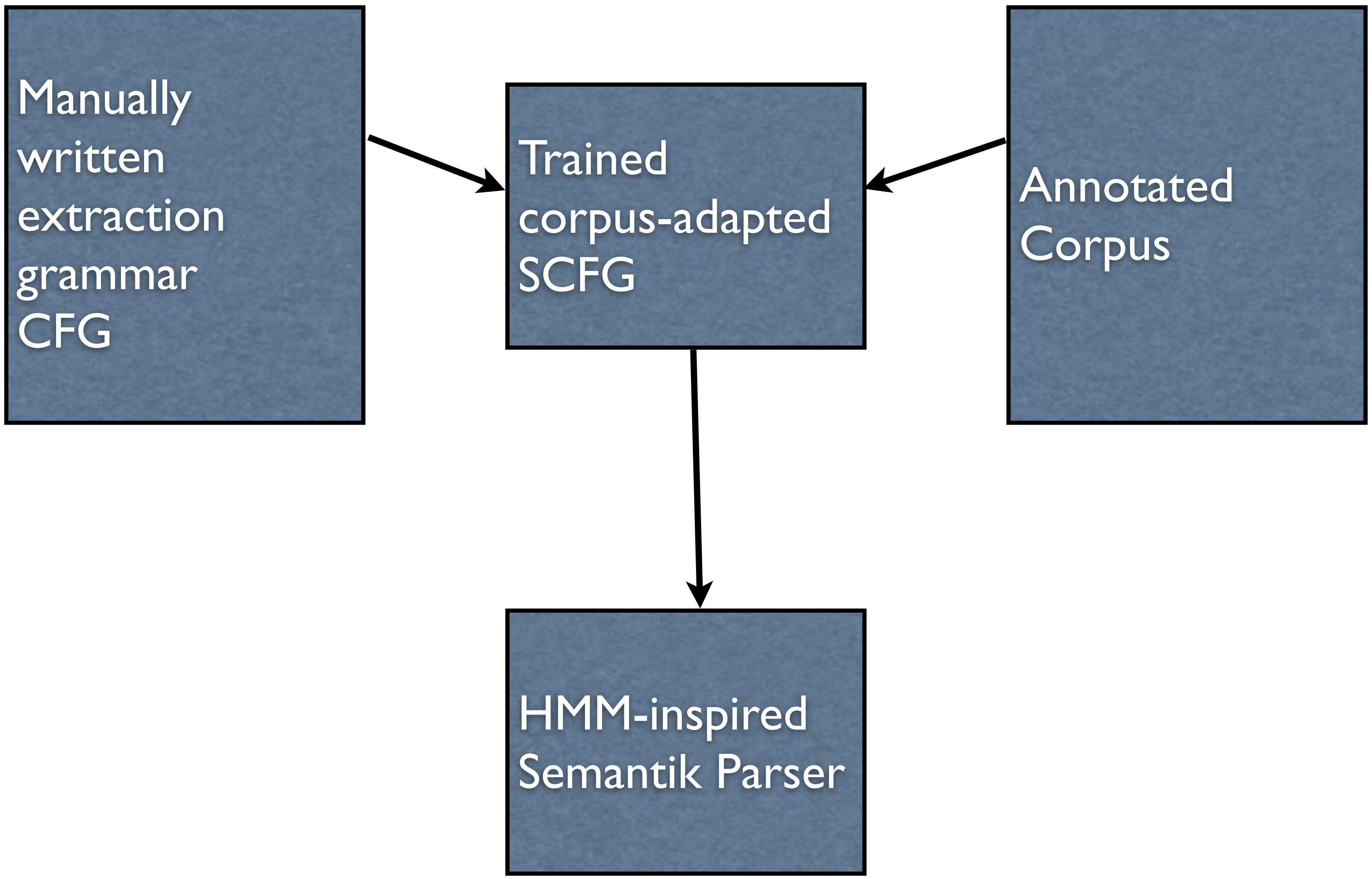
IE



Ontology population ontology learning

Projekt TheseusOrdo TechWatch

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

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

Collect statistics

P(Dr | **TLHonorific***)* = 1⁄5 (choice of one term among five equiprobable ones),
P(Dr | **NGFirstName***)* ≈ 1/*N*, where *N* is the number of all known words (untrained ngram behaviour).

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 :- ;

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(Dr | **TLHonorific***)* = 1/5 (choice of one term among five equiprobable ones),
P(Dr | **NGFirstName***)* ≈ 1/*N*, where *N* is the number of all known words (untrained ngram behaviour).

Parse corpus

Collect statistics

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

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