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

HumanMachine

Hybrid Language Processing
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)
```

<table>
<thead>
<tr>
<th>Unternehmen</th>
<th>Jahr</th>
<th>Größe</th>
<th>Betrag</th>
<th>Tendenz</th>
<th>Differenz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compaq</td>
<td>1998</td>
<td>Umsatz</td>
<td>31 Mrd. USD</td>
<td>+</td>
<td>27%</td>
</tr>
</tbody>
</table>
Example: news about turnover

turnover(Company, Year, Manner, Amount, Tendency, Difference)

<table>
<thead>
<tr>
<th>Unternehmen</th>
<th>Jahr</th>
<th>Größe</th>
<th>Betrag</th>
<th>Tendenz</th>
<th>Differenz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compaq</td>
<td>1998</td>
<td>Umsatz</td>
<td>31 Mrd. USD</td>
<td>+</td>
<td>27%</td>
</tr>
</tbody>
</table>

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.

Dienstag, 8. Februar 2011
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.

<table>
<thead>
<tr>
<th>Unternehmen</th>
<th>Jahr</th>
<th>Größe</th>
<th>Betrag</th>
<th>Tendenz</th>
<th>Differenz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compaq</td>
<td>1998</td>
<td>Umsatz</td>
<td>31 Mrd. USD</td>
<td>+</td>
<td>27%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unternehmen</th>
<th>Jahr</th>
<th>Größe</th>
<th>Betrag</th>
<th>Tendenz</th>
<th>Differenz</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>2003</td>
<td>Umsatz</td>
<td>21,6 Mrd. $</td>
<td>+</td>
<td>10 %</td>
</tr>
</tbody>
</table>

Dienstag, 8. Februar 2011
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.
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

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 - 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 - 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

Combiner

IE

IE

IE

IE
Multi-Strategy

Combiner

IE

IE

IE

IE
Example: NER
Example: NER

Problem:
- Ambiguities
- Bracketing

LOC 2
Wort1
Wort2
Wort3
Wort4
Wort5

LOC 3
ORG 4
LOC 5

Combiner

Ling Pipe
Open NLP
BiQue
Sprout

Dienstag, 8. Februar 2011
Example: NER

Problem:
- Ambiguities
- Bracketing

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

Dienstag, 8. Februar 2011

Good news:*

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

Example: NER

Problem:
- Ambiguities
- Bracketing

LingPipe
OpenNLP
BiQue
Sprout

Combiner

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

Example: Template Filling

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

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

MEM - Maximum Entropy Modeling
DOP - Data-Oriented Parsing


Dienstag, 8. Februar 2011
Example: Template Filling

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

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

Example: Template Filling

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

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

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

Feature based Strategies

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

Dienstag, 8. Februar 2011
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

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


Dienstag, 8. Februar 2011
Co-Training & Bootstrapping

Bootstrapper
Co-Training & Bootstrapping

Bootstrapper

Classifier 1

Classifier 2
Co-Training & Bootstrapping
Co-Training &
Bootstrapping

Bootstrapper

Classifier 1

Classifier 2

Dienstag, 8. Februar 2011
Co-Training & Bootstrapping

Bootstrapper

Classifier 1

Classifier 2

Dienstag, 8. Februar 2011
Co-Training & Bootstrapping

Bootstrapper

Classifier 1

Classifier 2
Co-Training & Bootstrapping
Co-Training & Bootstrapping

Initial data (seed)

Bootstrapper

Classifier 1

Classifier 2
Co-Training & Bootstrapping

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

Initial data (seed)

Bootstrapper

Classifier 1

Classifier 2
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

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.

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

Interaction of text classifier and pattern acquisition
• 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?  
  \[
  \text{lives}_\text{in}(\text{Town:}, \text{Pers:Bill Gates})
  \]

  What is a CEO?  
  \[
  \text{is}_\text{a}(\text{Pos:CEO,Conc:})
  \]

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


Dienstag, 8. Februar 2011
QA and Hybrid IE

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

Where does Bill Gates live?
\[ \text{lives_in(Town:?, Pers:Bill Gates)} \]

What is a CEO?
\[ \text{is_a(Pos:CEO, Conc:?)} \]

Domain open answering of definition questions from the Web

Was ist XYZ?

Problem: How to find optimal ranking of answer candidates?

Wikipedia as Blueprint!

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

What is XYZ?
Wikipedia as Blueprint!

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

What is XYZ?

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!

Web-QA

Unsupervised Learning of Feature Model

What is XYZ?

Prop
- automatic computation of POS and NEG training examples
- lex-sent feature-templates via dependency analysis
- Maximum Entropy Modeling

According to similarity of Wikipedia?

Dienstag, 8. Februar 2011
Wikipedia as Blueprint!

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

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.
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.
TEG - Tree Extraction Grammars

Manually written extraction grammar CFG

Trained corpus-adapted SCFG

HMM-inspired Semantik Parser

Annotated Corpus

Hand coded grammars

tenterm 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 :- ;
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.
Yesterday, <Person> Dr Simmons </Person>, the distinguished scientist presented the discovery.
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.

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

$P(Dr \mid NGFirstName) \approx \frac{1}{N}$, where $N$ is the number of all known words (untrained ngram behaviour).
Hand coded grammars
nonterm start Text;
concept Person;
ngram NGFirstName;
ngram NGLastName;
ngram NGNone;
termlist TLHonorific = Mr Mrs Miss Ms <2>Dr;
Person :- <2>TLHonorific NGLastName;
Text :- <11>NGNone Text;
Text :- <2>Person Text;
Text :- <2>;

Parse corpus

adapt rules

Collect statistics

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

\begin{align*}
P(Dr | TLHonorific) &= \frac{1}{5} \text{ (choice of one term among five equiprobable ones),} \\
P(Dr | NGFirstName) &\approx \frac{1}{N}, \text{ where } N \text{ is the number of all known words (untrained ngram behaviour).}
\end{align*}
TEG - Experiments

MUC-7 NER task

<table>
<thead>
<tr>
<th></th>
<th>HMM entity extractor</th>
<th>Emulation using TEG</th>
<th>DIAL Rules</th>
<th>Full TEG system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>R 86.91 P 85.13 F1 86.01</td>
<td>R 86.31 P 86.83 F1 86.57</td>
<td>R 81.32</td>
<td>R 93.75 P 87.53 F1 93.75</td>
</tr>
<tr>
<td>Org</td>
<td>R 87.94 P 89.75 F1 88.84</td>
<td>R 85.94 P 89.53 F1 87.7</td>
<td>R 82.74</td>
<td>R 93.36 P 88.05 F1 89.49</td>
</tr>
<tr>
<td>Location</td>
<td>R 86.12 P 87.2 F1 86.66</td>
<td>R 83.93 P 90.12 F1 86.91</td>
<td>R 91.46</td>
<td>R 89.53 P 90.49 F1 87.05</td>
</tr>
</tbody>
</table>

ACE-2 relation extraction

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

INC relation extraction

<table>
<thead>
<tr>
<th></th>
<th>Partial match results</th>
<th>Exact match results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Recall Prec F</td>
<td>Recall Prec F</td>
</tr>
<tr>
<td>PersonAffiliation</td>
<td>89.61 94.52 92.00</td>
<td>75.33 79.46 77.33</td>
</tr>
<tr>
<td>OrgLocation</td>
<td>85.32 77.78 80.00</td>
<td>76.47 72.22 74.29</td>
</tr>
<tr>
<td>Acquisition</td>
<td>76.00 86.36 80.85</td>
<td>68.00 77.27 72.34</td>
</tr>
</tbody>
</table>
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)
  • ...

Dienstag, 8. Februar 2011
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