Designing a Multi-Dimensional Space
for Hybrid Information Extraction (IE)

Christina Feilmayr, Klaudija Vojinovic, Birgit Pröll

Institute of Application Oriented Knowledge Processing, FAW
Overview
Overview

- Challenges in Information Extraction
- Motivating Hybrid Information Extraction (HybridIE)
- Fundamental Idea of Multi-Dimensional Space and HybridIE
- Scientific Findings, Project Modifications and Results
- Lessons Learned & Future Directions
Challenges in Information Extraction
Challenges in Information Extraction

- Common information extraction (IE) systems are imperfect
  - simple entity recognition: 90-98% correct results
  - template relation extraction: 50-60% correct results
Challenges in Information Extraction

• Common information extraction (IE) systems are imperfect
  — simple entity recognition: 90-98% correct results
  — template relation extraction: 50-60% correct results

• Developing an IE system is time- and labor intensive
Challenges in Information Extraction

• Common information extraction (IE) systems are imperfect
  — simple entity recognition: 90-98% correct results
  — template relation extraction: 50-60% correct results

• Developing an IE system is time- and labor intensive
  — KnowledgeBased (KB) IE: rules must be
Challenges in Information Extraction

• Common information extraction (IE) systems are imperfect
  — simple entity recognition: 90-98% correct results
  — template relation extraction: 50-60% correct results

• Developing an IE system is time- and labor intensive
  — Knowledge Based (KB) IE: rules must be
    ‣ sufficiently generic to extract the full extent of information
Challenges in Information Extraction

• Common information extraction (IE) systems are imperfect
  — simple entity recognition: 90-98% correct results
  — template relation extraction: 50-60% correct results

• Developing an IE system is time- and labor intensive
  — Knowledge-Based (KB) IE: rules must be
    ▸ sufficiently generic to extract the full extent of information
    ▸ sufficiently specific to extract relevant information
Challenges in Information Extraction

• Common information extraction (IE) systems are imperfect
  — simple entity recognition: 90-98% correct results
  — template relation extraction: 50-60% correct results

• Developing an IE system is time- and labor intensive
  — Knowledge Based (KB) IE: rules must be
    ‣ sufficiently generic to extract the full extent of information
    ‣ sufficiently specific to extract relevant information
  — Machine Learned (ML) IE: requires
Challenges in Information Extraction

• Common information extraction (IE) systems are imperfect
  — simple entity recognition: 90-98% correct results
  — template relation extraction: 50-60% correct results

• Developing an IE system is time- and labor intensive
  — KnowledgeBased (KB) IE: rules must be
    ‣ sufficiently generic to extract the full extent of information
    ‣ sufficiently specific to extract relevant information
  — MachineLearned (ML) IE: requires
    ‣ sufficiently large amount of training data
Challenges in Information Extraction

• Common information extraction (IE) systems are imperfect
  — simple entity recognition: 90-98% correct results
  — template relation extraction: 50-60% correct results

• Developing an IE system is time- and labor intensive
  — **KnowledgeBased** (KB) IE: rules must be
    ‣ **sufficiently generic** to extract the full extent of information
    ‣ **sufficiently specific** to extract relevant information
  — **MachineLearned** (ML) IE: requires
    ‣ **sufficiently large amount of training data**
    ‣ **appropriate set of features**
Motivating Hybrid Information Extraction
Motivating Hybrid Information Extraction

• Possible solution is to combine KB and ML - *hybrid IE, multi-strategy IE*
Motivating Hybrid Information Extraction

- Possible solution is to combine KB and ML - *hybrid IE, multi-strategy IE*

- Overall aim of research work
  - Developing **methods and processes that enables a more precise IE**
  - **Methodology for selecting appropriate hybrid IE methods**
Motivating Hybrid Information Extraction

- Possible solution is to combine KB and ML - *hybrid IE, multi-strategy IE*

- Overall aim of research work
  - Developing *methods and processes that enables a more precise IE*
  - *Methodology for selecting appropriate hybrid IE methods*

- Main Contributions
  - *Concepts for hybrid methods and processes*
  - *Decision support for selecting hybrid methods* (primarily *multi-dimensional space*, extended to *evaluation matrix*)
  - *Test framework* for two different application scenario (eRecruitment: analyzing a CV corpus, News: extracting data from Reuters corpus)
Multi-Dimensional Space
Multi-Dimensional Space

- Support IE system designers in selecting an appropriate method (ML, hybrid concept) for IE task
Multi-Dimensional Space

- Support IE system designers in selecting an appropriate method (ML, hybrid concept) for IE task

- Design of multi-dimensional space: three axes that indicates
  - **IE task**: NE, TE, TR, ST
  - **hybrid concept**: sequential extraction (SE), rule base extension (RB), knowledge base extension (KB)
  - **granularity of used features** (feature level)
Multi-Dimensional Space

- Support IE system designers in selecting an appropriate method (ML, hybrid concept) for IE task

- Design of multi-dimensional space: three axes that indicates
  - **IE task**: NE, TE, TR, ST
  - **hybrid concept**: sequential extraction (SE), rule base extension (RB), knowledge base extension (KB)
  - **granularity of used features** (feature level)
Multi-Dimensional Space

- Support IE system designers in selecting an appropriate method (ML, hybrid concept) for IE task

- Design of multi-dimensional space: three axes that indicates
  - **IE task**: NE, TE, TR, ST
  - **hybrid concept**: sequential extraction (SE), rule base extension (RB), knowledge base extension (KB)
  - **granularity of used features** (feature level)

- Results in a set of quintuples \([h, fl, t, m, x]\) (data points in space), e.g.,
Multi-Dimensional Space

- Support IE system designers in selecting an appropriate method (ML, hybrid concept) for IE task

- Design of multi-dimensional space: three axes that indicates
  - **IE task**: NE, TE, TR, ST
  - **hybrid concept**: sequential extraction (SE), rule base extension (RB), knowledge base extension (KB)
  - **granularity of used features** (feature level)

- Results in a set of quintuples \([h, f_l, t, m, x]\) (data points in space), e.g.,
  - \([\text{Sequential Extraction}, \text{Level2}, \text{TE}, \text{SVM}, 0.87]\)
Multi-Dimensional Space

- Support IE system designers in selecting an appropriate method (ML, hybrid concept) for IE task

- Design of multi-dimensional space: three axes that indicates
  - **IE task**: NE, TE, TR, ST
  - **hybrid concept**: sequential extraction (SE), rule base extension (RB), knowledge base extension (KB)
  - **granularity of used features** (feature level)

- Results in a set of quintuples \([h, f_1, t, m, x]\) (data points in space), e.g.,
  - [Sequential Extraction, Level2, TE, SVM, 0.87]
  - [Sequential Extraction, Level2, TE, k-NN, 0.64]
Multi-Dimensional Space

- Support IE system designers in selecting an appropriate method (ML, hybrid concept) for IE task

- Design of multi-dimensional space: three axes that indicates
  - **IE task**: NE, TE, TR, ST
  - **hybrid concept**: sequential extraction (SE), rule base extension (RB), knowledge base extension (KB)
  - **granularity of used features** (feature level)

- Results in a set of quintuples \([h, fl, t, m, x]\) (data points in space), e.g.,
  - [Sequential Extraction, Level2, TE, SVM, 0.87]
  - [Sequential Extraction, Level2, TE, k-NN, 0.64]
  - [Sequential Extraction, Level2, TE, CRF, 0.91]
Concepts of HybridIE
Concepts of HybridIE

• Sequential extraction (SE)
Concepts of HybridIE

• Sequential extraction (SE)

- Preprocessing
- Knowledge Based Information Extraction
- Feature resulting from KB method
- Global feature, resulting from lexical and syntactic analysis
- ML Based Information Extraction
- Training with different ML methods
- 10-fold-cross validation
- Results of Hybrid IE
Concepts of HybridIE

- **Sequential extraction (SE)**

- **Rule base extension (RB)**

![Diagram of HybridIE concepts]

- Preprocessing
- Knowledge Based Information Extraction
- ML Based Information Extraction
- Training with different ML methods
- 10-fold-cross validation
- Results of Hybrid IE
Concepts of HybridIE

- **Sequential extraction (SE)**
  - Preprocessing
  - Knowledge Based Information Extraction
  - Feature resulting from KB method
  - Global feature, resulting from lexical and syntactic analysis
  - ML Based Information Extraction
  - Training with different ML methods
  - 10-fold-cross validation
  - RESULTS of Hybrid IE

- **Rule base extension (RB)**
  - Labeled Documents → XML → Annotations
  - Global Feature Specific Feature
  - Unlabeled Documents
  - Feature Vector Generation
  - Hypothesis Building
  - Mining Task
  - Association Rules
Concepts of HybridIE
Concepts of HybridIE

- Knowledge base extension (KB)
Concepts of HybridIE

- Knowledge base extension (KB)
Intermediate Results of CV Extraction
Intermediate Results of CV Extraction

- Data was preprocessed using
Intermediate Results of CV Extraction

• Data was preprocessed using
  — rule-based system (provided by industrial partner JoinVision)
Intermediate Results of CV Extraction

• Data was preprocessed using
  — rule-based system (provided by industrial partner JoinVision)
  — GATE, which provides the lexical syntactic features (for ML), and its BatchLearner
Intermediate Results of CV Extraction

- Data was preprocessed using
  - rule-based system (provided by industrial partner JoinVision)
  - GATE, which provides the lexical syntactic features (for ML), and its BatchLearner
  - MALLET API (for CRF)
Intermediate Results of CV Extraction

- Data was preprocessed using
  - rule-based system (provided by industrial partner JoinVision)
  - GATE, which provides the lexical syntactic features (for ML), and its BatchLearner
  - MALLET API (for CRF)

### Results of KB

<table>
<thead>
<tr>
<th>IE TASK</th>
<th>PAUM</th>
<th>SVM</th>
<th>kNN</th>
<th>CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESULTS</td>
<td>P R F</td>
<td>P R F</td>
<td>P R F</td>
<td>P R F</td>
</tr>
<tr>
<td><strong>SECTION INDICATOR</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.81 0.65 0.72</td>
<td>0.76 0.70 0.72</td>
<td>0.32 0.27 0.29</td>
<td>0.78 0.72 0.79</td>
</tr>
<tr>
<td>Level 2</td>
<td>0.91 0.90 0.91</td>
<td>0.93 0.87 0.90</td>
<td>0.75 0.43 0.54</td>
<td>0.99 0.99 0.99</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.99 0.92 0.95</td>
<td>0.99 0.91 0.95</td>
<td>0.74 0.36 0.48</td>
<td>1.0 0.99 0.99</td>
</tr>
<tr>
<td><strong>PERSONS’ NAME</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.55 0.59 0.57</td>
<td>0.56 0.59 0.57</td>
<td>0.39 0.53 0.44</td>
<td>0.68 0.71 0.68</td>
</tr>
<tr>
<td>Level 2</td>
<td>0.94 0.78 0.85</td>
<td>0.96 0.80 0.87</td>
<td>0.82 0.64 0.71</td>
<td>0.98 1.0 0.99</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.98 0.80 0.88</td>
<td>1.00 0.81 0.89</td>
<td>0.98 0.82 0.89</td>
<td>1.0 1.0 1.0</td>
</tr>
<tr>
<td><strong>JOB TITLE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.52 0.42 0.46</td>
<td>0.58 0.43 0.49</td>
<td>0.24 0.14 0.17</td>
<td>0.64 0.66 0.65</td>
</tr>
<tr>
<td>Level 2</td>
<td>0.56 0.44 0.49</td>
<td>0.56 0.46 0.50</td>
<td>0.17 0.08 0.11</td>
<td>0.69 0.69 0.69</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.86 0.84 0.85</td>
<td>0.84 0.80 0.82</td>
<td>0.74 0.44 0.55</td>
<td>0.99 1.0 0.99</td>
</tr>
<tr>
<td><strong>ADDRESS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.57 0.50 0.51</td>
<td>0.50 0.51 0.50</td>
<td>0.54 0.43 0.47</td>
<td>0.64 0.59 0.61</td>
</tr>
<tr>
<td>Level 2</td>
<td>0.72 0.68 0.69</td>
<td>0.72 0.70 0.70</td>
<td>0.57 0.48 0.50</td>
<td>0.76 0.82 0.79</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.95 0.95 0.95</td>
<td>0.98 0.98 0.98</td>
<td>0.67 0.54 0.59</td>
<td>1.0 0.99 0.99</td>
</tr>
</tbody>
</table>
Evaluation Matrix for Hybrid IE
Evaluation Matrix for Hybrid IE
Evaluation Matrix for Hybrid IE

• Identification of criteria for analyzing ML methods with respect to their appropriateness for hybrid IE
  — evaluation matrix
  — support for user to identify appropriate ML methods for defined hybrid IE use case
Evaluation Matrix for Hybrid IE

- Identification of criteria for analyzing ML methods with respect to their appropriateness for hybrid IE
  - evaluation matrix
    - support for user to identify appropriate ML methods for defined hybrid IE use case
- Dynamic/static criteria (domain-dependent, -independent)
Evaluation Matrix for Hybrid IE

- Identification of criteria for analyzing ML methods with respect to their appropriateness for hybrid IE
  - evaluation matrix
  - support for user to identify appropriate ML methods for defined hybrid IE use case

- Dynamic/static criteria (domain-dependent, -independent)
  - Characterization of data set: size of data set, language of documents, balance of +/-
Evaluation Matrix for Hybrid IE

- Identification of criteria for analyzing ML methods with respect to their appropriateness for hybrid IE
  - evaluation matrix
  - support for user to identify appropriate ML methods for defined hybrid IE use case

- Dynamic/static criteria (domain-dependent, -independent)
  - Characterization of data set: size of data set, language of documents, balance of +/-
  - Characterization of ML method: kind of classification, impact of imbalanced data set, feature selection
Evaluation Matrix for Hybrid IE

• Identification of criteria for analyzing ML methods with respect to their appropriateness for hybrid IE
  — evaluation matrix
  — support for user to identify appropriate ML methods for defined hybrid IE use case

• Dynamic/static criteria (domain-dependent, -independent)
  — Characterization of data set: size of data set, language of documents, balance of +/-
  — Characterization of ML method: kind of classification, impact of imbalanced data set, feature selection
  — Fitness of ML method (i.r.t hybrid IE): single/multi class learning, correlations, identification/avoidance of errors
Semi-supervised Concepts for HybridIE
Semi-supervised Concepts for HybridIE

designed to improve ML methods
Semi-supervised Concepts for HybridIE

- Challenge of **imbalanced data set** → **sampler**
  - removing negative examples, duplication of positive example
  - random over-/undersampling, context (random) undersampler, WEKA sampler
    
    removeFrequentValues

**designed to improve ML methods**
Semi-supervised Concepts for HybridIE

- Challenge of **imbalanced data set** → **sampler**
  - removing negative examples, duplication of positive example
  - random over-/undersampling, context (random) undersampler, WEKA sampler \( \text{removeFrequentValues} \)

- Challenge of **insufficient amount of training data** → **semi-supervised ML**
  - self-training, co-training, active learning
Semi-supervised Concepts for HybridIE

• Challenge of imbalanced data set → sampler
  — removing negative examples, duplication of positive example
  — random over-/undersampling, context (random) undersampler, WEKA sampler
    removeFrequentValues

• Challenge of insufficient amount of training data → semi-supervised ML
  — self-training, co-training, active learning

• GATE-Plugin for semi-supervised learning and sampling
Semi-supervised Concepts for HybridIE

- Challenge of **imbalanced data set** ➞ **sampler**
  - removing negative examples, duplication of positive example
  - random over-/undersampling, context (random) undersampler, WEKA sampler
    - `removeFrequentValues`

- Challenge of **insufficient amount of training data** ➞ **semi-supervised ML**
  - self-training, co-training, active learning

- GATE-Plugin for semi-supervised learning and sampling

- Best results (+3-5%)
  - **Sampler**: context undersampler
  - **Semi-supervised approach**: self-training (SVM), co-training (PAUM, SVM), active learning (2x SVM)
Summarization of Project Results
Summarization of Project Results

• In general hybrid IE considerably performs better
Summarization of Project Results

• In general hybrid IE **considerably performs better**
  
  — KB+CRF best (GATE-Plugin for statistical methods)
Summarization of Project Results

- In general hybrid IE **considerably performs better**
  - KB+CRF best (GATE-Plugin for statistical methods)
  - Semi-supervised approaches and sampling supplementary improve hybrid IE results
Summarization of Project Results

• In general hybrid IE **considerably performs better**
  — KB+CRF best (GATE-Plugin for statistical methods)
  — Semi-supervised approaches and sampling supplementary improve hybrid IE results

• Hybrid IE provides a **correction of KB-annotated results**
Summarization of Project Results

• In general hybrid IE considerably performs better
  — KB+CRF best (GATE-Plugin for statistical methods)
  — Semi-supervised approaches and sampling supplementary improve hybrid IE results

• Hybrid IE provides a correction of KB-annotated results

... BUT ...
Summarization of Project Results

• In general hybrid IE **considerably performs better**
  — KB+CRF best (GATE-Plugin for statistical methods)
  — Semi-supervised approaches and sampling supplementary improve hybrid IE results

• Hybrid IE provides a **correction of KB-annotated results**

... **BUT** ...

• Selection of ML methods for hybrid IE is a **non-trivial task**
Summarization of Project Results

- In general hybrid IE **considerably performs better**
  - KB+CRF best (GATE-Plugin for statistical methods)
  - Semi-supervised approaches and sampling supplementary improve hybrid IE results
- Hybrid IE provides a **correction of KB-annotated results**

... **BUT** ...

- Selection of ML methods for hybrid IE is a **non-trivial task**
- There is **no standard solution**, which methods perform best in all hybrid IE use cases
Summarization of Project Results

- In general hybrid IE *considerably performs better*
  - KB+CRF best (GATE-Plugin for statistical methods)
  - Semi-supervised approaches and sampling supplementary improve hybrid IE results
- Hybrid IE provides a *correction of KB-annotated results*

... *BUT* ...

- Selection of ML methods for hybrid IE is a *non-trivial task*
- There is *no standard solution*, which methods perform best in all hybrid IE use cases
- Evaluation matrix is one possible support for IE system developer
Lessons Learned & Future Directions
Lessons Learned & Future Directions

- Identification of methods that overcome a specific IE challenge

  - main challenges of IE: ambiguity, imprecision, incompleteness, inconsistency, uncertainty and reliability
Lessons Learned & Future Directions

- Identification of methods that overcome a specific IE challenge
  - main challenges of IE: **ambiguity, imprecision, incompleteness, inconsistency, uncertainty** and reliability

- Challenges in case of **incompleteness**
  - incomplete/missing attribute-value pairs, incomplete/missing constraints and conditions
  - missing analysis of descriptive information (analysis of context information)
Lessons Learned & Future Directions

• Identification of methods that overcome a specific IE challenge
  — main challenges of IE: ambiguity, imprecision, incompleteness, inconsistency, uncertainty and reliability

• Challenges in case of incompleteness
  — incomplete/missing attribute-value pairs, incomplete/missing constraints and conditions
  — missing analysis of descriptive information (analysis of context information)

• Approach to overcome incompleteness
Lessons Learned & Future Directions

• Identification of methods that overcome a specific IE challenge
  — main challenges of IE: **ambiguity, imprecision, incompleteness, inconsistency, uncertainty and reliability**

• Challenges in case of **incompleteness**
  — incomplete/missing attribute-value pairs, incomplete/missing constraints and conditions
  — missing analysis of descriptive information (analysis of context information)

• Approach to overcome incompleteness
  — identification of incompleteness‘ characteristics
Lessons Learned & Future Directions

• Identification of methods that overcome a specific IE challenge
  — main challenges of IE: ambiguity, imprecision, incompleteness, inconsistency, uncertainty and reliability

• Challenges in case of incompleteness
  — incomplete/missing attribute-value pairs, incomplete/missing constraints and conditions
  — missing analysis of descriptive information (analysis of context information)

• Approach to overcome incompleteness
  — identification of incompleteness' characteristics
  — selection of methods (text-/data mining), which are appropriate to overcome incompleteness
Lessons Learned & Future Directions

• Identification of methods that overcome a specific IE challenge
  — main challenges of IE: ambiguity, imprecision, incompleteness, inconsistency, uncertainty and reliability

• Challenges in case of incompleteness
  — incomplete/missing attribute-value pairs, incomplete/missing constraints and conditions
  — missing analysis of descriptive information (analysis of context information)

• Approach to overcome incompleteness
  — identification of incompleteness‘ characteristics
  — selection of methods (text-/data mining), which are appropriate to overcome incompleteness
  — recommendation model (domain-dependent/independent)
Lessons Learned & Future Directions

• Identification of methods that overcome a specific IE challenge
  — main challenges of IE: ambiguity, imprecision, incompleteness, inconsistency, uncertainty and reliability

• Challenges in case of incompleteness
  — incomplete/missing attribute-value pairs, incomplete/missing constraints and conditions
  — missing analysis of descriptive information (analysis of context information)
Lessons Learned & Future Directions

• Identification of methods that overcome a specific IE challenge
  — main challenges of IE: **ambiguity, imprecision, incompleteness, inconsistency, uncertainty and reliability**

• Challenges in case of **incompleteness**
  — incomplete/missing attribute-value pairs, incomplete/missing constraints and conditions
  — missing analysis of descriptive information (analysis of context information)

Text Mining supported Information Extraction (TEMsIE)
... talk about „Characterization & Resolution of Incompleteness in (WWW) Information Extraction“
WebS2012 Workshop@DEXA (Sept., 05 2012, 10am)