

Designing a Multi-Dimensional Space

for Hybrid Information Extraction (IE)

Christina Feilmayr, Klaudija Vojinovic, Birgit Pröll

Institute of Application Oriented Knowledge Processing, FAW



BRIDGE Brückenschlagprogramm-2

Christina Feilmayr, September 04 2012

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
 - **KnowledgeBased** (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
 - **KnowledgeBased (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
 - **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

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

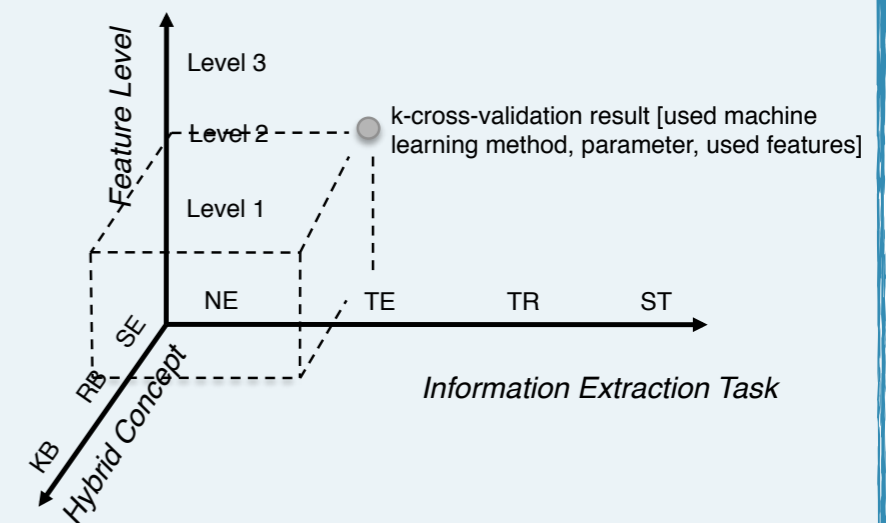
Multi-Dimensional Space

- 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

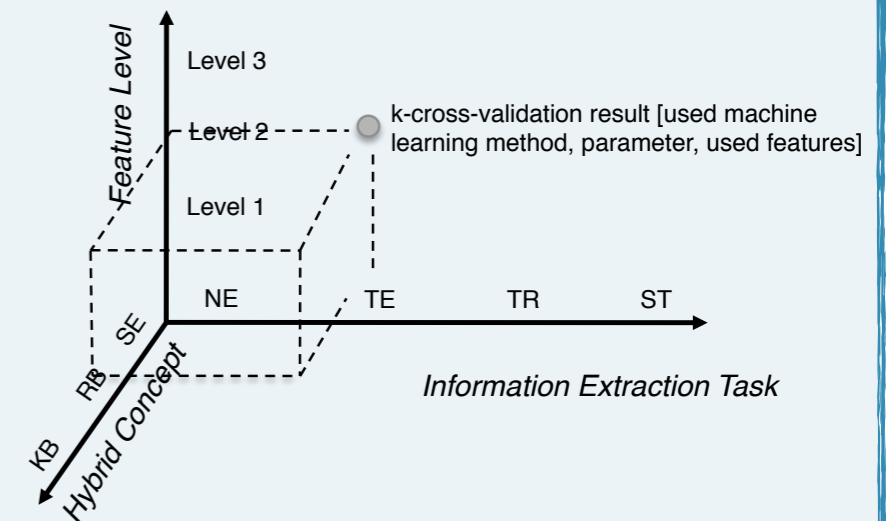
- Design of multi-dimensional space: three axes that indicates
 - IE task:** NE, TE, TR, ST
 - hybrid concept:** sequential extraction (SE), rule base extension (RB), knowledge base extension (KB)
 - granularity of used features** (feature level)



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)

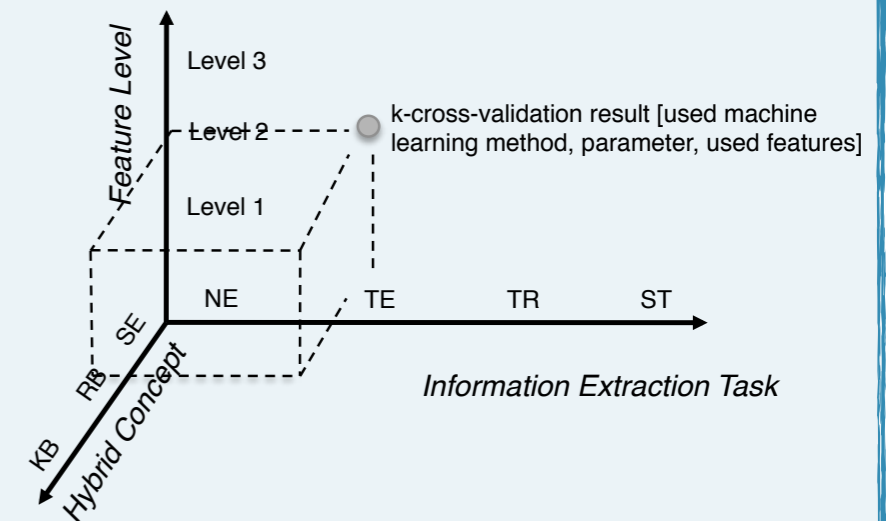


- Results in a set of quintuples $[h, fl, t, m, x]$ (data points in space), e.g.,

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)

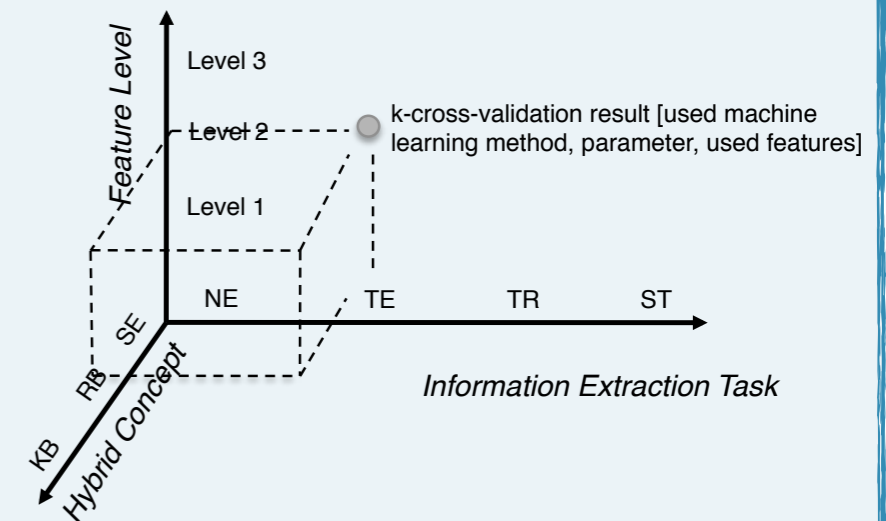


- Results in a set of quintuples $[h, fl, t, m, x]$ (data points in space), e.g.,
 - $[Sequential\ Extraction, Level2, TE, SVM, 0.87]$

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)

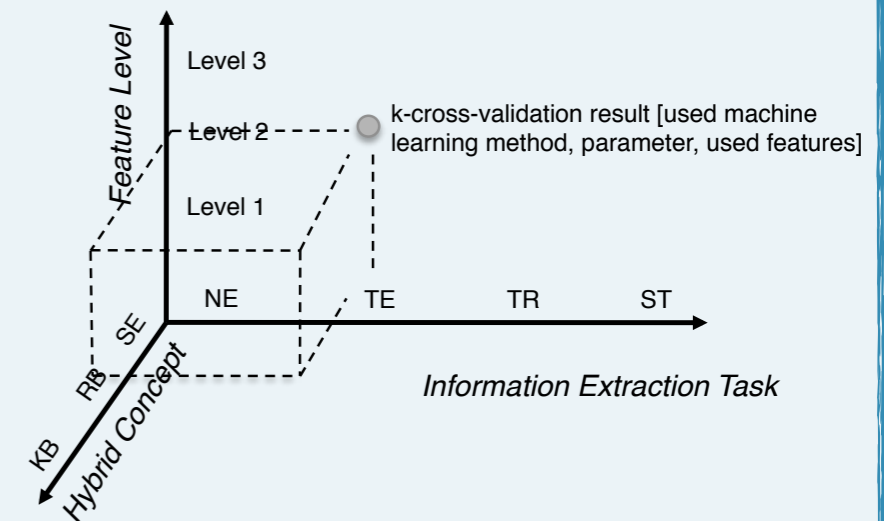


- Results in a set of quintuples $[h, fl, t, m, x]$ (data points in space), e.g.,
 - [Sequential Extraction, Level2, TE, SVM, 0.87]
 - [Sequential Extraction, Level2, TE, k-NN, 0.64]

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]

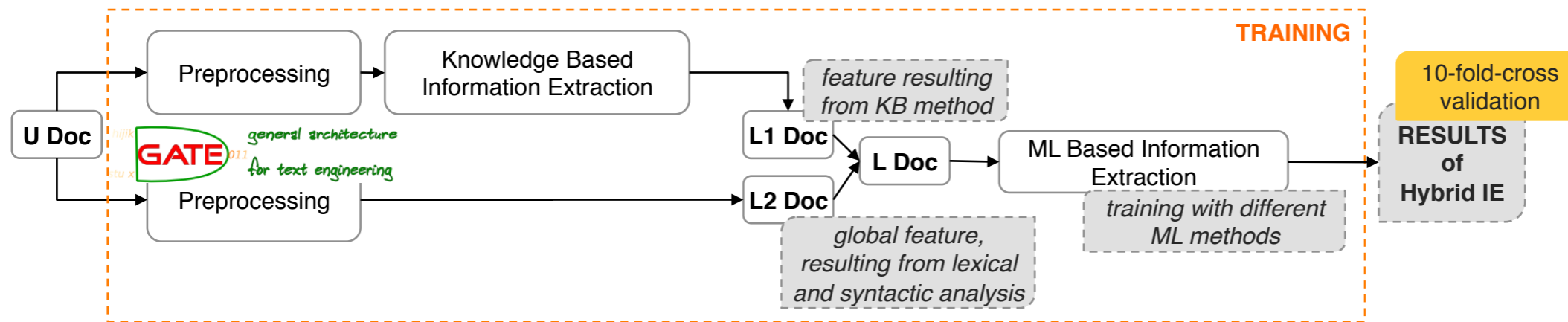
Concepts of HybridIE

Concepts of HybridIE

- **Sequential extraction (SE)**

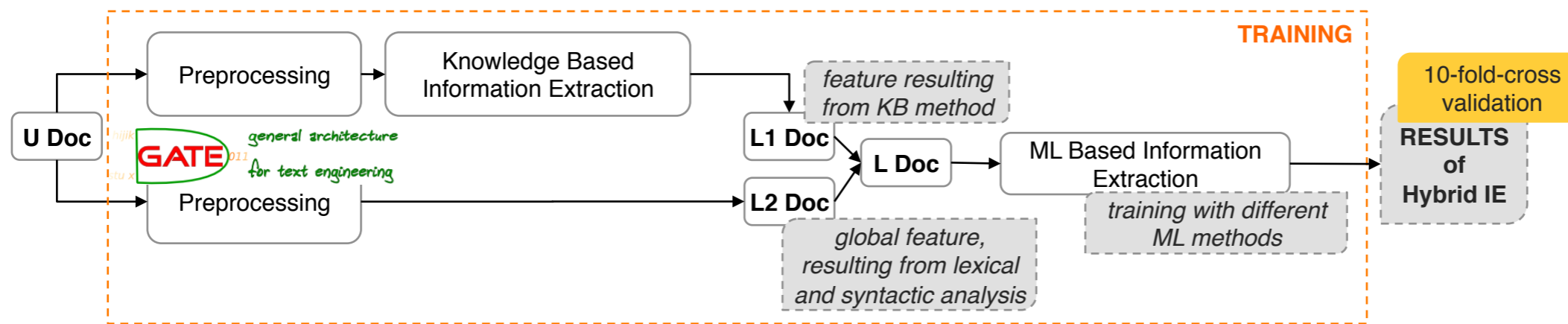
Concepts of HybridIE

- Sequential extraction (SE)



Concepts of HybridIE

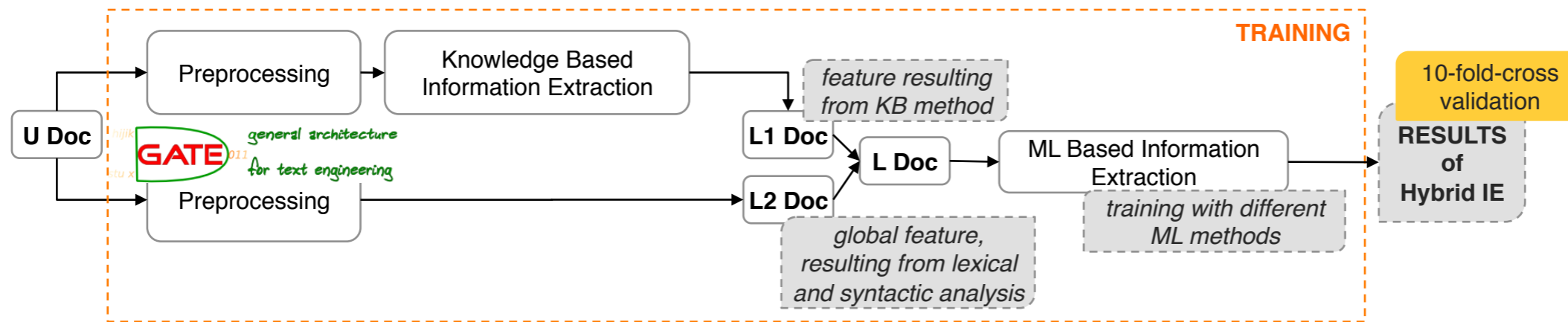
- Sequential extraction (SE)



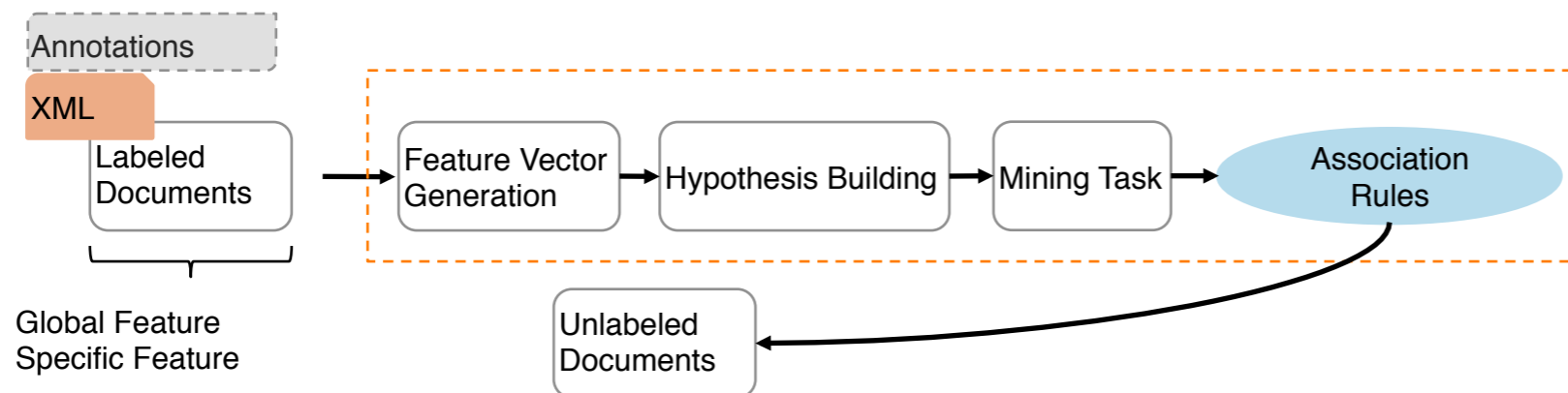
- Rule base extension (RB)

Concepts of HybridIE

- Sequential extraction (SE)



- Rule base extension (RB)



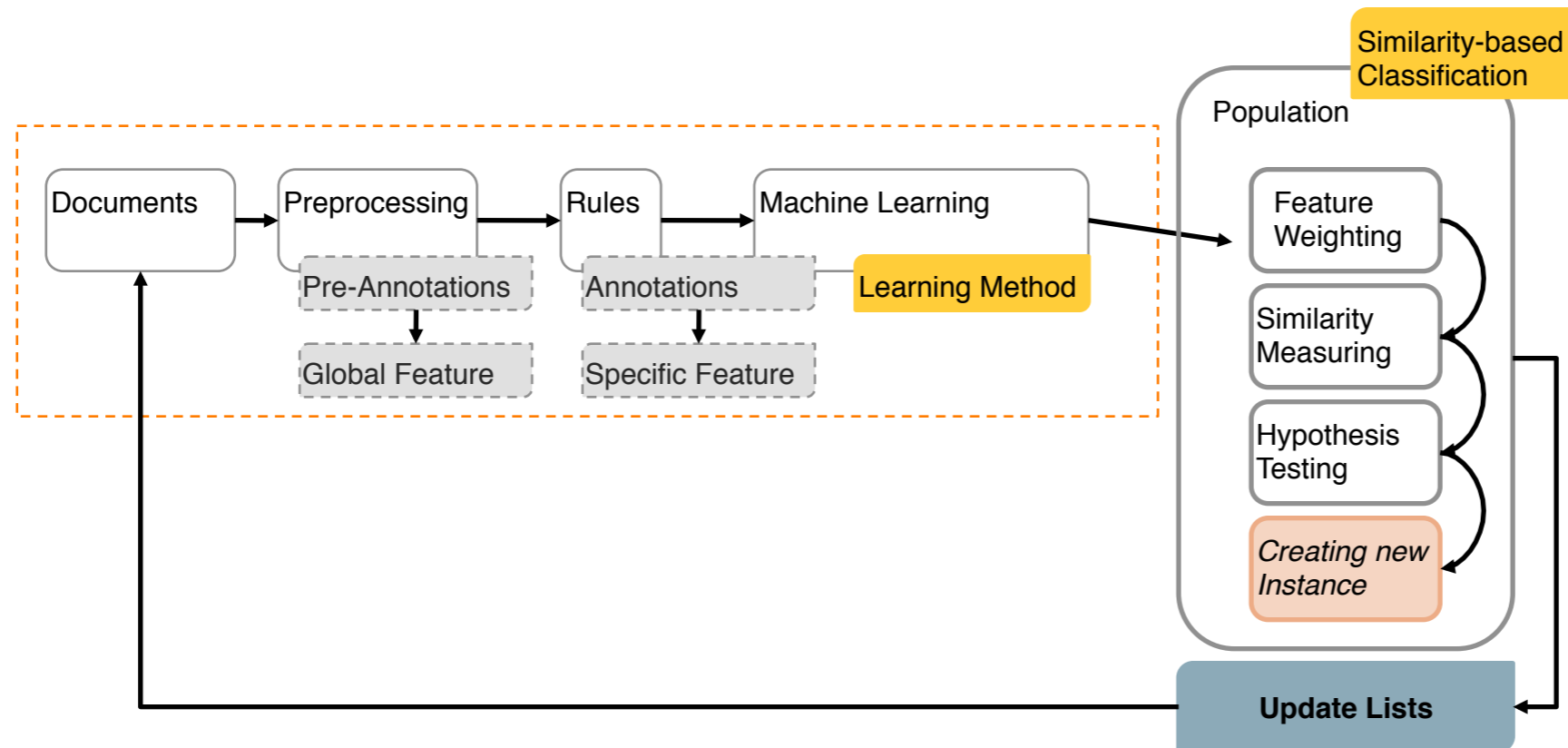
Concepts of HybridIE

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

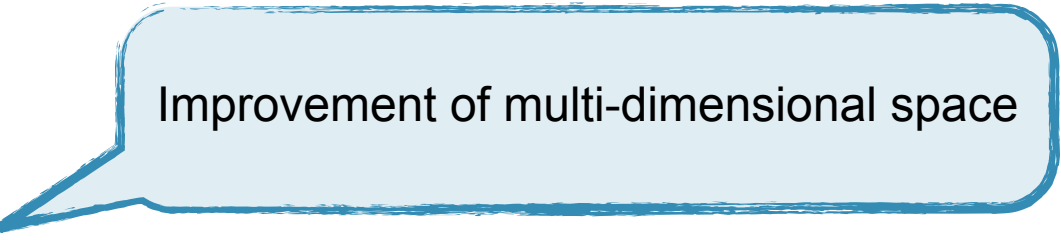
- Data was preprocessed using
 - rule-based system (provided by industrial partner JoinVision)
 - GATE**, which provides the lexical syntactic features (for ML), and its BatchLearner
 - MALLET** API (for CRF)

IE TASK		PAUM			SVM			kNN			CRF		
		P	R	F	P	R	F	P	R	F	P	R	F
SECTION INDICATOR	Results of KB	Precision (P) = 0.91 Recall (R) = 0.86 F1-measure (F) = 0.84											
	<i>Level 1</i>	0.81	0.65	0.72	0.76	0.70	0.72	0.32	0.27	0.29	0.78	0.62	0.69
	<i>Level 2</i>	0.91	0.90	0.91	0.93	0.87	0.90	0.75	0.43	0.54	0.99	0.99	0.99
	<i>Level 3</i>	0.99	0.92	0.95	0.99	0.91	0.95	0.74	0.36	0.48	1	0.99	0.99
PERSONS' NAME	Results of KB	Precision (P) = 0.86 Recall (R) = 0.82 F1-measure (F) = 0.84											
	<i>Level 1</i>	0.55	0.59	0.57	0.56	0.59	0.57	0.39	0.53	0.44	0.68	0.71	0.68
	<i>Level 2</i>	0.94	0.78	0.85	0.96	0.80	0.87	0.82	0.64	0.71	0.98	1	0.99
	<i>Level 3</i>	0.98	0.80	0.88	100	0.81	0.89	0.98	0.82	0.89	1	1	1
JOB TITLE	Results of KB	Precision (P) = 0.93 Recall (R) = 0.94 F1-measure (F) = 0.93											
	<i>Level 1</i>	0.52	0.42	0.46	0.58	0.43	0.49	0.24	0.14	0.17	0.64	0.66	0.65
	<i>Level 2</i>	0.56	0.44	0.49	0.56	0.46	0.50	0.17	0.08	0.11	0.69	0.69	0.69
	<i>Level 3</i>	0.86	0.84	0.85	0.84	0.80	0.82	0.74	0.44	0.55	0.99	1	0.99
ADDRESS	Results of KB	Precision (P) = 0.86 Recall (R) = 0.75 F1-measure (F) = 0.79											
	<i>Level 1</i>	0.57	0.50	0.51	0.50	0.51	0.50	0.54	0.43	0.47	0.64	0.59	0.61
	<i>Level 2</i>	0.72	0.68	0.69	0.72	0.70	0.70	0.57	0.48	0.50	0.76	0.82	0.79
	<i>Level 3</i>	0.95	0.95	0.95	0.98	0.98	0.98	0.67	0.54	0.59	1	0.99	0.99

Evaluation Matrix for Hybrid IE

Evaluation Matrix for Hybrid IE

Improvement of multi-dimensional space



Improvement of multi-dimensional space

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



Improvement of multi-dimensional space

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)



Improvement of multi-dimensional space

Evaluation Matrix for Hybrid IE

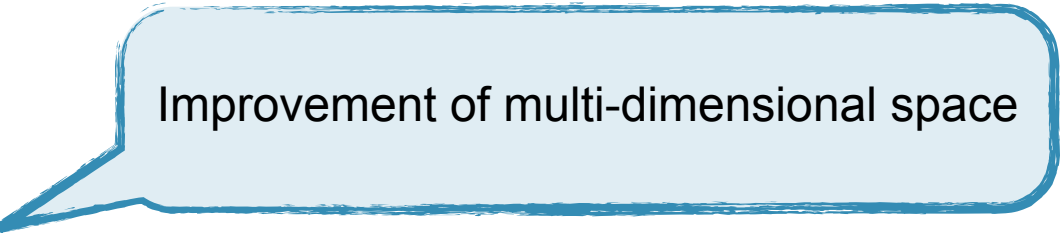
- Identification of criteria for analyzing ML methods with respect to their appropriateness for hybrid IE
 - **evaluation matrix**
 - support for user to identify appropriate ML methods for defined hybrid IE use case
- Dynamic/static criteria (domain-dependent, -independent)
 - Characterization of data set: **size of data set, language of documents, balance of +/-**



Improvement of multi-dimensional space

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**



Improvement of multi-dimensional space

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

designed to improve ML methods

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

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
- Challenge of **insufficient amount of training data** → **semi-supervised ML**
 - self-training, co-training, active learning
- GATE-Plugin for semi-supervised learning and sampling

1
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
- 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)



Christina Feilmayr

Johannes Kepler University Linz | AUSTRIA

cfeilmayr@faw.jku.at

<http://www.faw.jku.at>