UNIVERSITÄT LEIPZIG

Increasing Recall of Process Model Matching by Improved Activity Label Matching

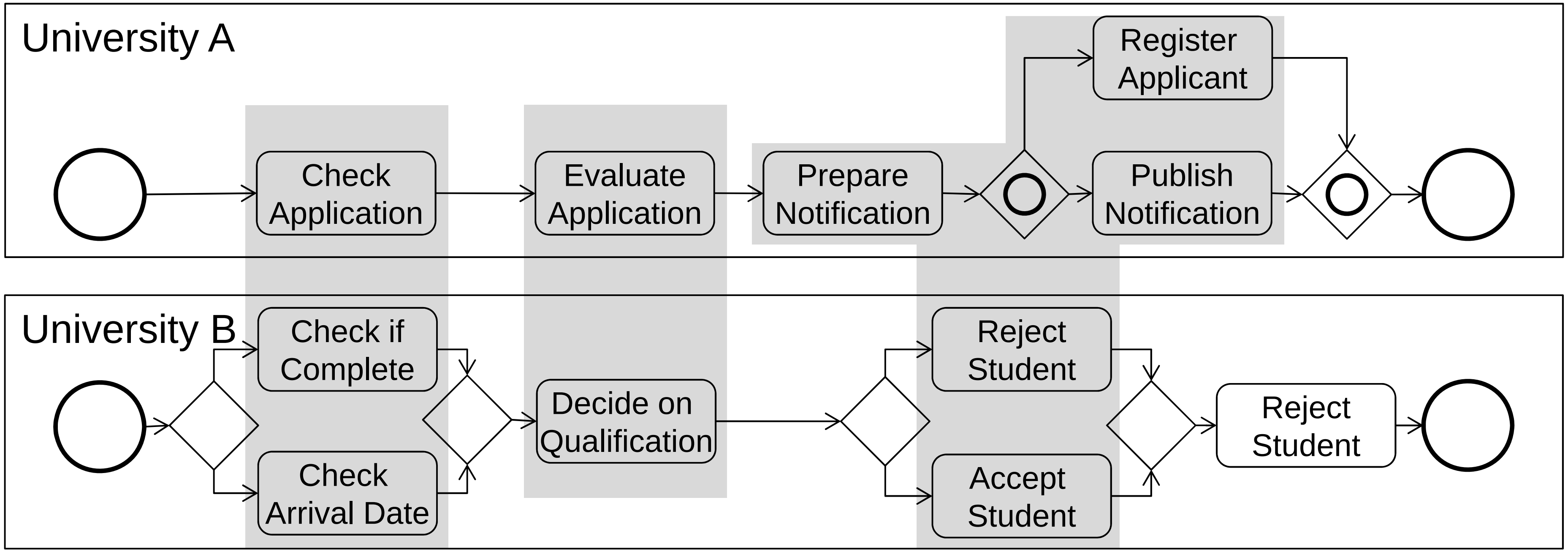
Christopher Klinkmüller, Ingo Weber, Jan Mendling, Henrik Leopold & André Ludwig

University of Leipzig
Endowed Chair of Logistics Information Systems
klinkmueller@wifa.uni-leipzig.de

Agenda

- Motivation
- Techniques
  - Basic Bag-of-Words Similarity
  - Bag-of-Words Similarity with Label Pruning
- Analysis
  - Evaluation Results
  - Matching Challenges
- Summary

Motivation



- Existing approaches yield high precision, but low recall
- High recall necessary to be a useful tool

➢ Goal: increasing recall without sacrificing precision

Basic Process Matching Algorithm

1. Determine Similarity Matrix
   - Calculate **sim(a_1,a_2)** for all (a_1,a_2) with $a_1 \in P_1$ and $a_2 \in P_2$
2. Select Matches
   - Define threshold **t**
   - Propose all activity pairs with sim(a_1,a_2) > t

- Similarity Measures
  - Basic Bag-of-Words Similarity
  - Bag-of-Words Similarity with Label Pruning

Bag-of-Words Similarity

[illegible]

send email to apply

[illegible]

Bag-of-Words Similarity with Label Pruning

Step 4: ignore words from larger bag

Step 4: consider words from larger bag with

1. highest term frequency (tf) (prune$_{max}$)
2. highest tf in model collection (prune$_{all}$)
3. highest tf in model pair (prune$_{2p}$)

			check	documents
	sim = 0.717		1	1

tf = 14
tf = 12
	check	sim = 0.8		

tf = 4
tf = 7
	documents	sim = 0.775		

tf = 9
tf = 1
	complete			

tf = 3
tf = 6
	time			

Evaluation Setup

- Model collection and gold standard (Leopold et al., 2012)
  - 9 models → 36 pairs
  - 1:1 activity matches for each pair
- Measures
  - For each model pair: Precision, Recall & F_1-Value
  - For model collection: Mean of Precision, Recall and F_1-Value
- Used Word Similarities
  - lev: Levenshtein Distance (Levenshtein, 1966)
  - lin: Semantic notion based on WordNet (Lin, 1998)
  - max: maximum lev and lin
  - s.lev:
  - s.lin: } words are stemmed using jwi-WordNet-stemmer[1]
  - s.max:
- Threshold Sampling
  - Over the intervall: [0..1]
  - In steps of size 0.05

[1] http://projects.csail.mit.edu/jwi

Evaluation Results

Results for Basic Bag-of-Words

Results for Pruning

Prototype			Precision	Recall	F_1
Markov			0.421	0.263	0.315
ICoP			0.506	0.255	0.294
basic	max	0.75	0.748	0.299	0.363
basic	s.lev	0.75	0.808	0.304	0.372
$\text{prune}_{\text{max}}$	s.lev	0.75	0.735	0.331	0.393
$\text{prune}_{\text{coll}}$	s.lin	0.70	0.468	0.450	0.409
prune_{2p}	s.lev	0.80	0.689	0.356	0.407

Challenge Analysis

- Based on best result
  - 912 matches (223 TP, 381 FP, 308 FN)
- Approach
  - Three Researchers involved
  - Manual challenge clustering
  - Resolution of differences in discussions

Challenge Categories

1. Label specificity
   – Refers to the granularity of the labels
2. Wording challenges
   – Refers to the words of the label
3. Term semantics
   – Problems may arise from the meaning of the words
4. Process structure
   – Control flow characteristics that lead to wrong decisions

Challenge Analysis

class	challenge	#	FP+FN	TP
1	detail of information	463	0.86	0.14
3	compound words	412	0.85	0.15
1	implicit objects	290	0.86	0.14
2	different conditions	249	0.92	0.08
1	higher-level activity	223	0.76	0.24
3	semantic relation	136	1.00	0.00
4	control flow position	120	1.00	0.00
1	action/object combinations	99	0.83	0.17
4	different roles	75	1.00	0.00
4	case differentiation	59	0.73	0.27
2	abbreviations	27	0.93	0.07
2	domain specificity	25	0.96	0.04
3	spelling errors	21	0.86	0.14
2	sentence structure	17	0.77	0.23
2	inverse	9	1.00	0.00

Summary

- Contribution
  - Improvements by neglecting label structure and label pruning
  - Identification of matching problems and their importance
  - Code will be made available in September on Google Code http://source.google.com/p/jpmmt
- Future work
  - Extending evaluation base
  - Solving most important issues
    - Detail of information
    - Compound Words
    - Implicit Objects