Automated public data refining

Martin Lipták
Ing. Ján Suchal

14.3.2012

Introduction

- What are public data?
  - registers of companies
  - registers of organizations
  - public procurements
  - public contracts
  - lists of debtors
- What about problems with public data?
  - mistypings, duplicates, disambiguities
  - it is really messy

Public data (mess) example

Mgr. Juraj Široký MBA	Strmý vŕšok 8137/137 Bratislava - Záhorská Bystrica
Ing. Juraj Široký	Strmý vŕšok 137 Bratislava
Ing. Juraj Široký	Priekopnícka 30 Bratislava
Juraj Široký	Gromové 28 Praha Česká republika

Field normalization

- squeeze white spaces
- convert to lower cases
- transliterate language-specific chars (č, ť, á)
- extract academic degrees to separate fields
- extract address parts to separate fields

String similarity for typos

- Levenshtein (edit) distance
  - minimum number of edit operations needed to transform one string into another
  - John Dough <=> John Doe (3)
- N-gram similarity
  - break string into n-gram sets (n can be 2, 3, 4, 5, ...)
  - |A intersection B| / |A union B| (jaccard similarity)
  - example
    - John Doe => (__J, _Jo, Joh, ohn, hn_, n_D, _Do, Doe, oe_, e__)
    - John Dough => (__J, _Jo, Joh, ohn, hn_, n_D, _Do, Dou, oug, ugh, gh_, h__)
    - 7 / 15 = 51.42%

Heuristics based on relations

- e.g. occurrences in companies

Putting it all together

- machine learning
- preliminary experiment...

Preliminary experiment

- supervised machine learning
- logistic regression classifier

Data set

- slovak business register (foaf.sk)
- detecting duplicates in people table
- 4,298 out of 569,999 records
  - name
  - address
- no company relations
  - companies and occurrences tables
- training and testing on all possible pairs
  - $4,298 * (4,298 + 1)/2 = 9,238,551$ samples
- label
  - our baseline is current foaf.sk duplicate detection

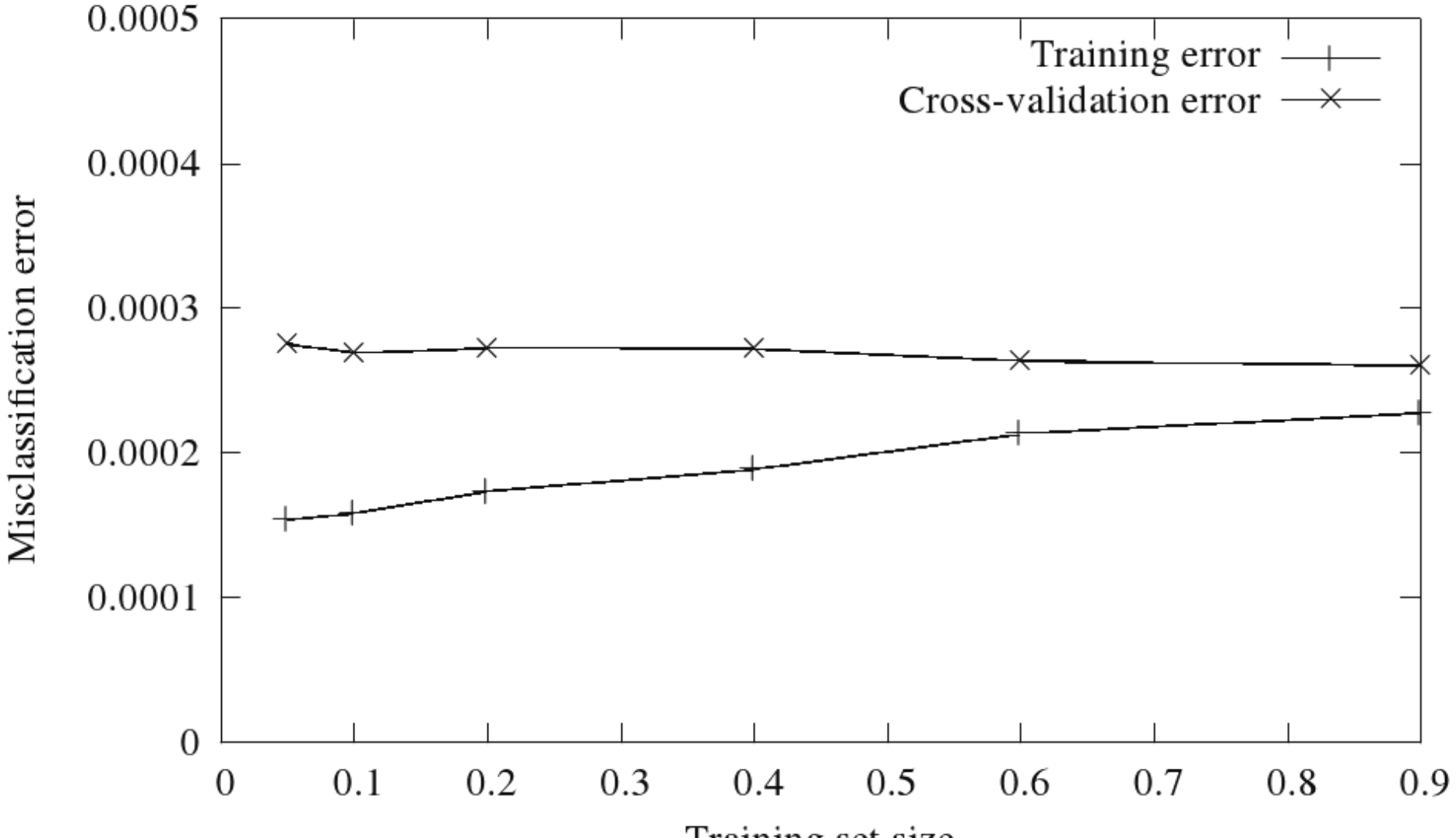
Features

- equal names
- equal addresses
- levenshtein distance of names
- levenshtein distance of addresses
- n-gram similarity of names
- n-gram similarity of addresses
- combination of academic degrees
  - feature for every possible pair of occurring degrees
  - testing compatibility of degrees
- disjunction of academic degrees
  - degree occurring in one of two compared samples

Results

Feature set	FP	FN	Precision	Recall	F_1 score
=(labels)	0	0	1	1	1
=(names), =(addresses)	142	13	0.8777	0.9874	0.9293
L(names), L(addresses)	326	3	0.8782	0.9923	0.9318
2G(names), 2G(addresses)	142	13	0.8777	0.9874	0.9293
3G(names), 3G(addresses)	142	13	0.8777	0.9874	0.9293
4G(names), 4G(addresses)	142	13	0.8777	0.9874	0.9293
5G(names), 5G(addresses)	142	13	0.8777	0.9874	0.9293
6G(names), 6G(addresses)	142	13	0.8777	0.9874	0.9293
2G(names + degrees), 2G(addresses)	138	40	0.8779	0.9612	0.9177
3G(names + degrees), 3G(addresses)	138	46	0.8772	0.9554	0.9147
4G(names + degrees), 4G(addresses)	136	50	0.8784	0.9516	0.9135
5G(names + degrees), 5G(addresses)	135	53	0.8788	0.9486	0.9124
6G(names + degrees), 6G(addresses)	135	54	0.8787	0.9477	0.9119
L(names), L(addresses), degree combinations	135	39	0.8803	0.9622	0.9194
L(names), L(addresses), degree disjunctions	135	23	0.882	0.9777	0.9274

Learning curve



Future work

- data set with our own labels (instead of foaf.sk baseline labels)
- inclusion of relations to companies in features (currently we use only names, addresses and degrees)
- address normalization, distance of address coordinates
- performance
  - levenshtein automata, min-hashing
  - vowpal wabbit algorithm
  - sliding block comparing

Thank you :)