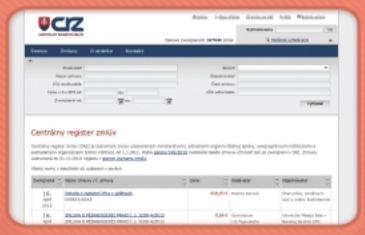
Automated Public Data Refining

Register of companies

Register of organizations

Contracts



Lists of debtors

Cultural grants

Public procurements

Municipality statistics

Database

Internet domains

Business bulletins

Problems

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

- include relations between compared entities
- e.g. common occurrences in organizations

Methods

- Manually tuned parameters
- Machine learning (supervised, clustering)
- Blocking methods
- User assistance

Experiment 1

- supervised machine learning
- logistic regression classifier
- training and testing on all possible pairs of data set entities

Data set

- 4,298 entities from register of companies (data provided by foaf.sk)
- baseline label
  - current foaf.sk duplicate detection
- attributes
  - name
  - address

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	F_1 score
=(labels)	0	0	1
=(names), =(addresses)	142	13	0.9293
L(names), L(addresses)	326	3	0.9318
2G(names), 2G(addresses)	142	13	0.9293
3G(names), 3G(addresses)	142	13	0.9293
4G(names), 4G(addresses)	142	13	0.9293
5G(names), 5G(addresses)	142	13	0.9293
6G(names), 6G(addresses)	142	13	0.9293
2G(names + degrees), 2G(addresses)	138	40	0.9177
3G(names + degrees), 3G(addresses)	138	46	0.9147
4G(names + degrees), 4G(addresses)	136	50	0.9135
5G(names + degrees), 5G(addresses)	135	53	0.9124
6G(names + degrees), 6G(addresses)	135	54	0.9119
L(names), L(addresses), degree combinations	135	39	0.9194
L(names), L(addresses), degree disjunctions	135	23	0.9274

Conclusion

- developed our evaluation framework
- trained feature weights can be reused by other duplicate detection systems
- machine learning can be used for public data refining

Experiment 2

- new data set
- new features
- support vector machine classifier

Data set

- ~ 250 entities from register of organizations
- 3 hours of work (selection and manual labeling)
- attributes
  - name
  - address
- relations
  - occurrence in a common organization

Features

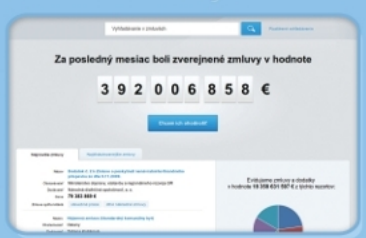
- address-related features broken into separate features for address parts (street, number1, number2, code, town)
- binary feature for occurrence in a common organization
- binary feature for presence of blank address

Preliminary conclusion

- breaking address into parts improves results
- inclusion of heuristics improves results
- degree features don't work due to small data set
- SVM yields better results than logistic regression
- classifier

Foaf.sk

Otvorenezmluvy.sk



Statistiky-domen.sk

Naseobce.sk

Martin Lipták, supervised by Ing. Ján Suchal