Stemming and Search Strategies for East European Language

Ljiljana Dolamic, Jacques Savoy
Computer Science Department
University of Neuchatel, Switzerland
www.unine.ch/info/clef/

East European Languages

- Hungarian
- Slavic Languages
  - Bulgarian
  - Czech
  - Russian

Hungarian

- Ob-Ugric language
- Large number of cases

Hungarian

- Stem – plural – possesion – case
  - gyereke-i-nke-t

    child –Pl – PlPoss – Acc
- Derivatinals
  - jelent – és (meaning)

    to mean – der

Hungarian

- Compound constructions

hétvégé = hét + vég

weekend = week/seven+ end

Savoy,J. Report on CLEF-2003 monolingual tracks: Fusion of probabilistic models for effective monolingual retrieval

Bulgarian

- Southern Slavic Language
- Cyrillic
- No cases
- Definite article

Bulgarian

- stem – plural – artical

  - вечер – и – те
    evening – PL – the
  - геро(й) – Ø – ят/я
    hero – Ø – the
  - слаб – а – та
    weak – f,sg – the

Bulgarian

- Derivationals

  - Българ – СК – и – те

    stem - der - PL – the

    (the Bulgarian)

Problems with Bulgarian

- Mutation of –Я–
  - бял – белота
    (white - whiteness)
  - грях – грехове
    (sin - sins)
- Elision of vowel
  –Е– or –Ъ–
  - орел – орли
    (eagle - eagles)
  - топъл –топла
    (warm, m - f)
- Palatalisation
  1. К,Г,Х ➔ Ч,Ж,Ш
     - око – очи
       (eye – eyes)
     - Бог – Боже
       (God, Nom - Voc)
  2. К,Г,Х ➔ Ц,З,С
     - вълк – вълци
       (wolf - wolves)

Czech

- Western Slavic Language
- Seven case system
- stem – case
  - pán – ovi

    sir (N,L,sg)
  - mlad – ou

    young(A,sg,f)

Czech

- Stem - case

case / gendre	nominative	dative singulier	dative plural
masculine	pán (sir)	pán - ovi	pán - ům
feminine	žen – a (woman)	žen - ě	žen - ám
neutre	mlad – é (young)	mlad - ému	mlad - ým

Czech

- Derivationals
  - klavír – ist – a (pianist)
    piano – der – case
  - Žid – ovk – a (Jewish woman)
    Jew – der – case

Problems with Czech

- Fleeting – E –
  - zámek – zámkem
    (castel, Nom – Ins)
  - otec – otcův
    (father – father's)
- ů → o
  - stůl – stoly
    (table – tables)
- Consonant softening
  - matka – matčin
    (mother – mother's)
  - drahý – drazí
    (dear, Nom, sg – pl)
  - mokrý – mokří
    (wet, Nom, sg – pl)
  - český – čeští
    (Czech, adj,
    Nom, sg – pl)

Russian

- Eastern Slavic Language
- Cyrillic
- Six cases
- stem – case
  - книг – а
    book (N, sg)
  - хорош – ая
    good (N, sg, f)

Evaluation

- *4-grams*
- IR models
  - Okapi
  - DFR
  - LM
  - *tf·idf*

Evaluation Hungarian

Model Q=TD	word	dec	4-grams	
Okapi	0.3231	0.3629	0.3445	
DFR IneC2	0.3525	0.3897	0.3527	
LM(λ=0.35)	0.3118	0.3482	0.3153	
tf·idf	0.2344	0.2532	0.2345	

Evaluation Hungarian

Model Q=TD	word	dec	4-grams	jmorph*
Okapi	0.3231	0.3629	0.3445	0.3509
DFR IneC2	0.3525	0.3897	0.3527	0.3480
LM(λ=0.35)	0.3118	0.3482	0.3153	0.3155
tf·idf	0.2344	0.2532	0.2345	0.2224

***jmorh – Java port for hunmorph morphological avalyzer (http://mokk.bme.hu/resouces/ir)**

Evaluation Hungarian

Model Q=TD	word	dec	4-grams	jmorph
Okapi	0.3231*	0.3629*	0.3445	**0.3509**
DFR IneC2	**0.3525**	**0.3897**	**0.3527**	0.3480
LM(λ=0.35)	0.3118*	0.3482*	0.3153	0.3155*
tf·idf	0.2344*	0.2532*	0.2345	0.2224*

Evaluation Bulgarian

Model Q=TD		light	deriv.	4-grams
Okapi		0.3155	0.3425	0.3022
DFR IneC2		0.3423	0.3606	0.3156
LM(λ=0.35)		0.3175	0.3368	0.2868
tf·idf		0.2103	0.2143	0.2105

Evaluation Bulgarian

Model Q=TD	word	light	deriv.	4-grams
Okapi	0.2035	0.3155	0.3425	0.3022
DFR IneC2	0.2215	0.3423	0.3606	0.3156
LM(λ=0.35)	0.2083	0.3175	0.3368	0.2868
tf·idf	0.1636	0.2103	0.2143	0.2105
Over light	-32.8%	baseline	+5.8%	-5.9%

Evaluation Bulgarian

Model Q=TD	word	light	deriv.	4-grams
Okapi	0.2035*	0.3155*	0.3425*	0.3022
DFR IneC2	**0.2215**	**0.3423**	**0.3606**	**0.3156**
LM(λ=0.35)	0.2083*	0.3175*	0.3368*	0.2868*
tf·idf	0.1636*	0.2103*	0.2143*	0.2105*

Evaluation Czech

Model Q=TD	light		deriv.	4-grams
Okapi	0.3355		0.3255	0.3401
DFR GL2	0.3437		0.3342	0.3365
DFR IneC2	0.3539		0.3473	0.3517
LM(λ=0.35)	0.3263		0.3109	0.3204
tf·idf	0.2050		0.1984	0.2126

Evaluation Czech

Model Q=TD	light	light noAccent	deriv.	4-grams
Okapi	0.3355	0.3306	0.3255	0.3401
DFR GL2	0.3437	0.3359	0.3342	0.3365
DFR IneC2	0.3539	0.3473	0.3473	0.3517
LM(λ=0.35)	0.3263	0.3174	0.3109	0.3204
tf·idf	0.2050	0.2078	0.1984	0.2126

Evaluation Czech

Model Q=TD	light	light noAccent	deriv.	4-grams
Okapi	0.3355	0.3306*	0.3255*	0.3401*
DFR GL2	0.3437	0.3359	0.3342	0.3365
DFR IneC2	***0.3539***	***0.3473***	***0.3473***	***0.3517***
LM(λ=0.35)	0.3263*	0.3174*	0.3109*	0.3204*
tf·idf	0.2050*	0.2078*	0.1984*	0.2126*

Evaluation Russian

Model Q=TD	light	4-grams	
Okapi	0.1630	0.0917	
DFR GL2	0.1639	0.1264	
DFR InB2	0.1775	0.1052	
LM(λ=0.35)	0.1511	0.1246	
tf·idf	0.1188	0.0918	

Evaluation Russian

Model Q=TD	light	4-grams	snowball*
Okapi	0.1630	0.0917	0.1617
DFR GL2	0.1639	0.1264	0.1689
DFR InB2	0.1775	0.1052	0.1749
LM(λ=0.35)	0.1511	0.1246	0.1524
tf˙idf	0.1188	0.0918	0.1194

* http://snowball.tartarus.org/

Evaluation Russian

Model Q=TD	light	4-grams	snowball
Okapi	0.1630	0.0917*	0.1617
DFR GL2	0.1639	**0.1264**	0.1689
DFR InB2	**0.1775**	0.1052	**0.1749**
LM(λ=0.35)	0.1511*	0.1246	0.1524
tf·idf	0.1188*	0.0918*	0.1194*

Query-by-Query

- Hard topics
  - map < 0.1

Query-by-Query - Hungarian

- #411, #426, #436, #439, #446
- #436 , 'VIP divorces'
  - 0.0003 (DFR GL2, dec)

<title>**VIP válások**</title>
<desc>**Keressünk cikkeket híres emberek válásáról.**</desc>

  - VIP – df=0

Query-by-Query - Bulgarian

- light, 4 grams
  - #407
  - #412
  - #417
  - #422
  - #428
  - #429
  - #435
- agressive
  - #412
  - #417
  - #422
  - #428
  - #435

Query-by-Query - Bulgarian

- #429, 'Water Health Risks '

<title>**Рискове за здравето, причинени от вода**</title>
<desc>**Намерете документи, които съдържат информация за рисковете за здравето от замърсена или заразена вода.**</desc>

		deriv.	light
Q	здравето	здрав	здрав
D	здравен	здрав	здравн
D	здравна		
D	здравното		
Q	заразена	зараг	заразн
D	заразата		зараг

Query-by-Query - Czech

- #411, #422, #428, #430, #435, #439, #446

- #430, 'Cosmetic procedures'
  - 0.0025 (tf·idf, Q=TDN, 4grams)
  - 0.1553 (DFR GL2, Q=D, light)

- #411, 'Best picture Oscar '
  - 0.0053 (DFR GL2, Q=TDN, light)

<title>**Oskar za nejlepší film**</title>
<desc>**Jaký titul získal v březnu 2002 Oskara za nejlepší film?**</desc>

Query-by-Query - Russian

- 4 grams
  - #176
  - #180
  - #185
  - #186
  - #189
  - #192
  - #194
  - #196
  - #198
- light
  - #176
  - #185
  - #186
  - #189
  - #192
  - #195
  - #196

Query-by-Query - Russian

- #192, ‘System change and family planning in East Germany ’
  - 0.0034 (DFR InB2,light, Q=TDN)

<title>**Трансформация и семейное планирование в Восточной Германии**</title>

<desc>**Найти документы, в которых описываются тенденции в области деторождения и семейное планирование в Восточной Германии после объединения.**</desc>

  - 1 relevant item

Query-by-Query - Russian

- #171, ‘Sibling relations ’
  - 0.0089(DFR InB2, light, Q=TDN)

<title>**Отношения между родными братьями и сестрами**</title>
<desc>**Найдите документы, которые подробно описывают развитие отношений между родными сестрами и братьями.**</desc>

  - 2 relevant items
  - семейиые – family

Conclusion

- Is stemming effective?
- Best performing retrieval model
- Hard topics

Stemming and Search Strategies for East European Languages

Ljiljana Dolamic,Jacques Savoy
Computer Science Department
University of Neuchatel, Switzerland
www.unine.ch/info/clef/