University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming

Craig Macdonald
Vassilis Plachouras
Ben He
Iadh Ounis

{craigm,vassilis,ben,ounis}@dcs.gla.ac.uk

UNIVERSITY of GLASGOW

Objectives

- Test per-field normalisation as a technique for the combination of multi-lingual web evidence

- Investigate per-language stemming in multilingual Web IR

Per-language Stemming

- Three indices:
  - Apply no stemming at all
  - Apply Porters English stemmer to entire corpus
  - Apply correct stemmer for language of each document
- For languages-specific, apply Snowball stemmers, except:
  - English: Porters English stemmer
  - Icelandic: Danish Snowball Stemmer
  - Hungarian: Hunspell/Hunstem
  - Greek: No stemmer

Language Identification

- Identify one language for each document from:
  - Language classifier data
  - URL
  - Metadata
  - Expected languages in that domain

- Our Information Retrieval Platform
  - Written in Java
  - Core version available for download from http://ir.dcs.gla.ac.uk/terrier/
- We used a version supporting UTF-8 encoding
- Detect encoding of each document in EuroGOV
  - Based on language, HTTP and META metadata

Web Retrieval with Terrier

- Three fields: Content, TITLE & Anchor Text
- Normalise each field w.r.t. field length individually before combining
  - New weighting scheme which is a refinement and derivative of the PL2 weighting scheme.
- URL Scoring
  - Favour documents with shorter URLs

Runs

	Index	Topic Metadata		Notes
uog SelStem	All Stemmers	x	x	Language classifier run on topics
uog AllStem	All Stemmers	Language	Target Domain	
uog AllStem NP	All Stemmers	Language	Target Domain	
uog PorStem	Porter Stemmer	x	Target Domain	
uog NoStem	No Stemming	x	Target Domain	

Runs

- uogSelStem
  - No topic metadata used
  - Use language classifier to stem topics correctly
  - Use all stemmers index
- uogAllStem / uogAllStemNP
  - Use all stemmers index
  - Topic Language and target page domain metadata
- uogPorStem
  - Use Porter stemmed index
  - Target page domain metadata
- UogNoStemNLP
  - Use non-stemmed index
  - Target page domain metadata

Results (I)

(MRR) / Topics	*uog SelStem*	**uog NoStem NLP**	uog PorStem	uog AllStem	uog AllStem NP
All	0.4683	**0.5135**	0.5107	0.4827	0.4828

Results (II)

(MRR) / Topics	*uog SelStem*	**uog NoStem NLP**	uog PorStem	uog AllStem	uog AllStem NP
All	0.4863	**0.5135**	0.5107	0.4827	0.4828
DA	0.5168	0.5246	0.5098	**0.5857**	0.5829
DE	0.4467	0.4414	0.4567	**0.4780**	0.4689
EL	0.2047	0.3704	0.3659	0.3586	**0.4003**
EN	0.4988	**0.5578**	0.5240	0.5188	0.5239
ES	0.4198	0.4571	0.4635	0.4602	**0.4647**
FR	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
HU	0.2713	**0.5422**	**0.5422**	0.1142	0.1003
IS	0.3400	0.3222	0.3222	**0.3400**	**0.3400**
NL	0.6362	0.6226	**0.6551**	0.6444	0.6447
PT	0.5262	**0.5565**	0.5336	0.5048	0.5028
RU	0.4838	0.4724	**0.4975**	0.4838	0.4625

Results (III)

(MRR) / Topics	uog SelStem	**uog NoStem NLP**	uog PorStem	uog AllStem	uog AllStem NP
All	0.4863	**0.5135**	0.5107	0.4827	0.4828
NP only	0.4803	**0.5353**	0.5232	0.4952	0.4956
HP only	0.4531	0.4862	**0.4949**	0.4669	0.4666

Conclusions

- Language-specific stemming works
  - But requires robust and effective language identification
  - And stemmers!
- No stemming seems to be safe option
- [Jansen & Spink 2005] investigated a European search engine:
  - German – 1.9 terms
  - Spanish – 2.6 terms
- WebCLEF queries were much longer
  - 3.3 terms, 6.3 terms for German & Spanish respectively
  - Representative of real European user tasks?

Thank You

UNIVERSITY
of
GLASGOW