human language technology
center of excellence

Alternative Methods for Tokenization in Ad Hoc Search

Paul McNamee

18 September 2008

JOHNS HOPKINS
UNIVERSITY

Talk Outline

- Introduction
- Retrospective Experiments 2000-2007
- Tokenization Alternatives in 13 languages
  - words
  - stemmed words (Snowball)
  - character n-grams (n=4,5)
  - n-gram stems
  - automatically segmented words (Morfessor algorithm)
- TEL Results
- Persian Results
  - skipgrams (n-grams with skips)
- Summary

human language technology
center of excellence

CLEF Ad Hoc Test Sets (2000 – 2007)

	#docs	size	00	01	02	03	04	05	06	07	
Bulgarian (BG)	69 k	213 MB						49	50	50	149
Czech (CS)	82 k	178 MB								50	50
Dutch (NL)	190 k	540 MB		50	50	56					156
English (EN)	170 k	580 MB	33	47	42	54	42	50	49	50	367
Finnish (FI)	55 k	137 MB			30	45	45				120
French (FR)	178 k	470 MB	34	49	50	52	49	50	49		333
German (DE)	295 k	660 MB	37	49	50	56					192
Hungarian (HU)	50 k	105 MB						50	48	50	148
Italian (IT)	157 k	363 MB	34	47	49	51					181
Portuguese (PT)	107 k	340 MB					46	50	50		146
Russian (RU)	17 k	68 MB				28	34				62
Spanish (ES)	453 k	1086 MB		49	50	57					156
Swedish (SV)	143 k	352 MB			49	53					102

Ad Hoc Experimental Setup

- Only Title + Description runs
- No Relevance Feedback
- Separately indexed collections for each document set
  - Note: the document collections grew in some languages during certain years
- MAP measured using only the queries with at least one relevant document
- JHU HAIRCUT system
  - Language modeling approach: Ponte & Croft, SIGIR-98
  - Linear interpolation smoothing with parameter $\alpha = 0.5$ for all terms, for all tokenization types, in all languages.

Stemming

- **Applicable to alphabetic languages**
- **An approximation to lemmatization**
- **Principle: chop off affixes**
- **Improves recall in a Boolean system**
- **Used for Dutch, English, Finnish, French, German, Italian, Spanish, and Swedish**
- **Snowball rulesets also exist for Hungarian and Portuguese**

Most stemmers are rule-based

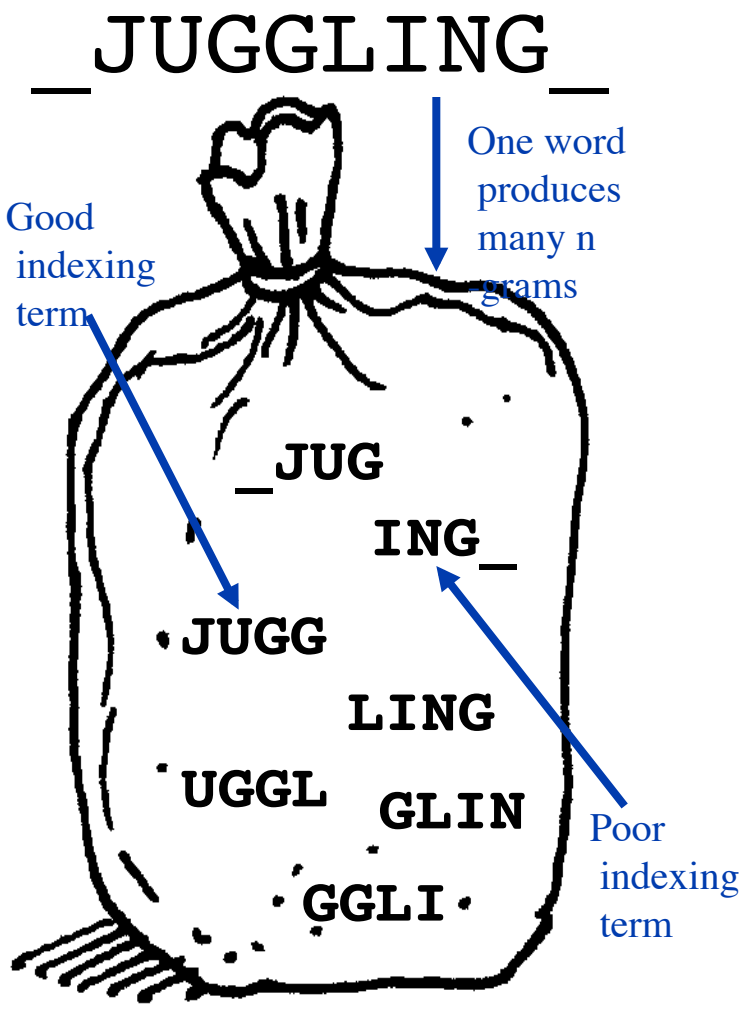
-ing => ε	juggling => juggl
-es => ε	juggles => juggl
-le => -l	juggle => juggl

The Snowball project provides high quality, rule-based stemmers for many European languages

http://snowball.tartarus.org/

human language technology
center of excellence

N-Gram Tokenization



- **Characterize text by overlapping sequences of *n* consecutive characters**
- **In alphabetic languages, *n* is typically 4 or 5**
- **N-grams are a language-neutral representation**
- **N-gram tokenization incurs both speed and disk usage penalties:**

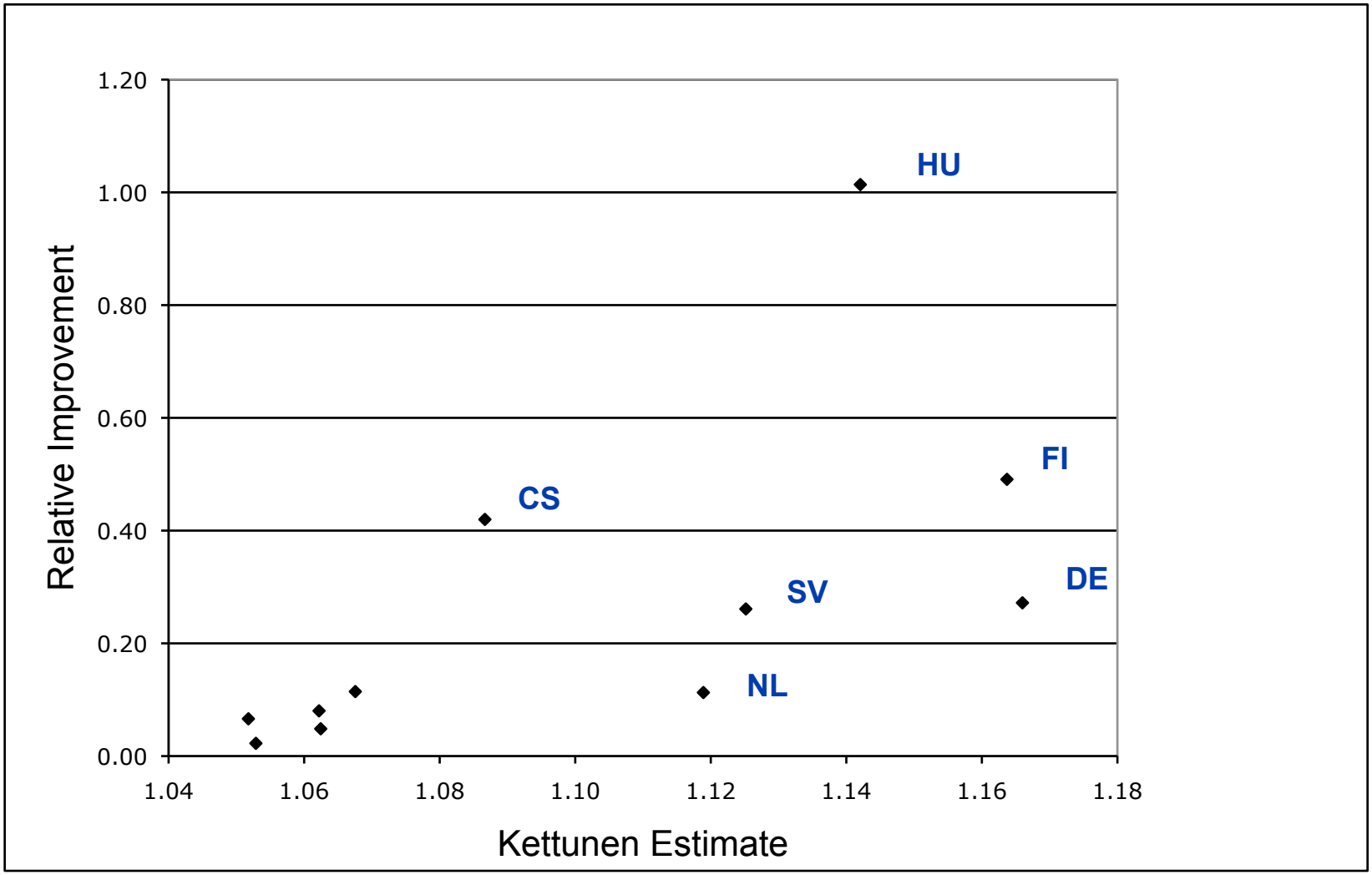
"Every character begins an n-gram"

Words vs. Stems vs. N-grams

	# topics	words	stems	4-grams	5-grams
BG	149	0.2164		0.3105	0.2820
CS	50	0.2270		0.3294	0.3223
DE	192	0.3303	0.3695	0.4098	0.4201
EN	367	0.4060	0.4373	0.3990	0.4152
ES	156	0.4396	0.4846	0.4597	0.4609
FI	120	0.3406	0.4296	0.4989	0.5078
FR	333	0.3638	0.4019	0.3844	0.3930
HU	148	0.1520		0.3192	0.3061
IT	181	0.3749	0.4178	0.3738	0.3997
NL	156	0.3813	0.4003	0.4219	0.4243
PT	146	0.3162		0.3358	0.3524
RU	62	0.2671		0.3406	0.3330
SV	102	0.3387	0.3756	0.4236	0.4271
Average		0.3195		0.3851	0.3880
Avg (8)		0.3719	0.4146	0.4214	0.4310

human language technology
center of excellence

5-grams and Morphological Complexity



JOHNS HOPKINS
UNIVERSITY

Least Common N-gram Stemming

- **Traditional (rule-based) stemming attempts to remove the morphologically variable portion of words**
  - **Negative effects from over- and under-conflation**

Hungarian	Bulgarian
_hun (20547)	_bul (10222)
hung (4329)	bulg (963)
unga (1773)	ulga (1955)
ngar (1194)	lgar (1480)
gari (2477)	gari (2477)
aria (11036)	aria (11036)
rian (18485)	rian (18485)
ian_ (49777)	ian_ (49777)

Short n-grams covering affixes occur frequently - those around the morpheme tend to occur less often. This motivates the following approach:

(1) For each word choose the least frequently occurring character 4 -gram (using a 4-gram index)

(2) Benefits of n-grams with run -time efficiency of stemming

Continues work in Mayfield and McNamee, 'Single N-gram Stemming', SIGIR 2003

Examples

Lang.	Word	Snowball	LC4
English	juggle	juggl	jugg
English	juggles	juggl	jugg
English	juggler	juggler	jugg
English	juggled	juggl	jugg
English	juggling	juggl	jugg
English	juggernaut	juggernaut	rnau
English	warred	war	warr
English	warren	warren	warr
English	warrens	warren	rens
English	warrant	warrant	warr
English	warring	war	warr

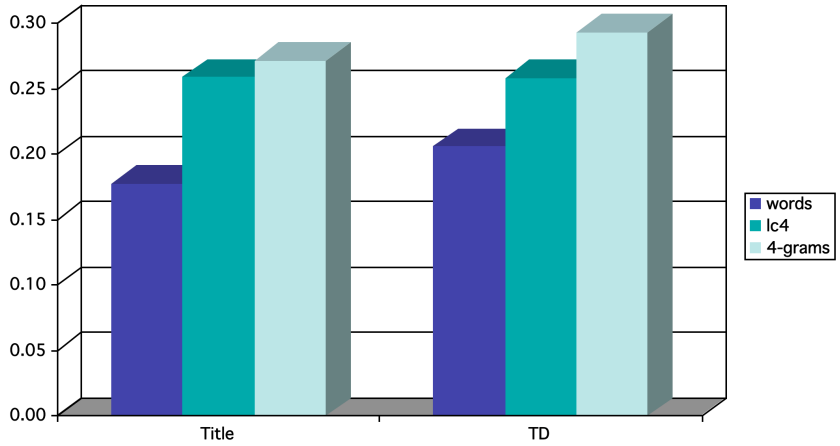
Lang.	Word	Snowball	LC4
Swedish	kontroll	kontroll	ntro
Swedish	kontrollerar	kontroller	ntro
Swedish	kontrollerade	kontroller	ntro
Swedish	kontrolleras	kontroller	ntro
English	pantry	pantri	antr
English	tantrum	tantrum	antr
English	marinade	marinad	inad
English	marinated	marin	rina
English	marine	marin	rine
English	vegetation	veget	etat
English	vegetables	veget	etab

All approaches to conflation, including no conflation at all, make errors.

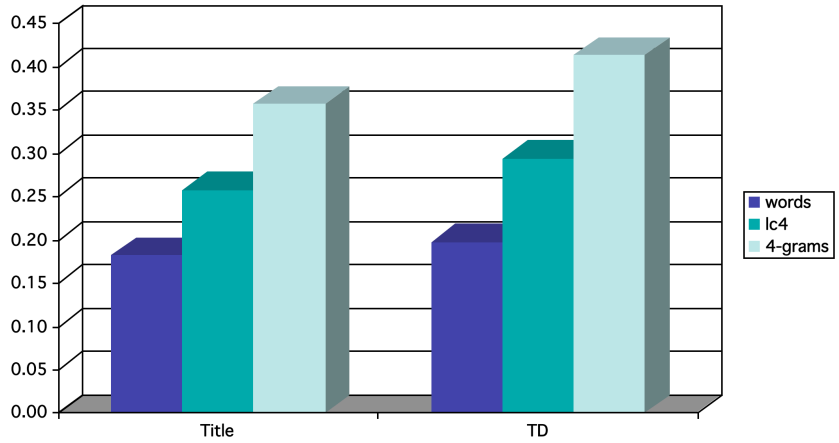
human language technology
center of excellence

N-gram Stemming Effectiveness

Bulgarian



Hungarian



CLEF 2005 data

- **4-grams dominate words**
  - **25-50% advantage in Bulgarian**
  - **Improvements even larger in Hungarian**
- **LC4 also dominates words**

Morfessor Segmentation

- **Minimum Description Length principle**
  - Given input wordlist: maximize model size + model fit
- **Produces segmentation of each word**
  - 70% of world languages have concatenative morphology
  - Only segments – does not transform letters in surface form
- **Examples**
  - affectionate = affect+ion+ate
  - unconcerned = un+concern+ed
- **For IR experiments added all segments to inverted file**
  - Extremely common 'terms' ignored (> 20% of docs), but term weighting controls high frequency terms impact

JOHNS HOPKINS
UNIVERSITY

IR experiments also reported in McNamee et al., "Don't Have a Stemmer? Be Un+concern+ed", SIGIR 2008 poster

Sample Tokenizations

Word	Snowball	Morfessor	5-grams
authored	author	author+ed	_auth, autho, uthor, thore, hored, ored_
authorized	author	author+ized	_auth, autho, uthor, thori, horiz, orize, rized, ized_
authorship	authorship	author+ship	_auth, autho, uthor, thors, horsh, orshi, rship, ship_
reauthorization	reauthor	re+author+ization	_reau, reaut, eauth, autho, uthor, thori, horiz, oriza, rizat, izati, zatio, ation, tion_
afoot	afoot	a+foot	_afoo, afoot, foot_
footballs	footbal	football+s	_foot, footb, ootba, otbal, tbaall, balls, alls_
footloose	footloos	foot+loose	_foot, footl, ootlo, otloo, tloos, loose, oose_
footprint	footprint	foot+print	_foot, footp, ootpr, otpri, tprin, print, rint_
feet	feet	feet	_feet, feet_
juggle	juggl	juggle	_jugg, juggl, uggle, ggle_
juggled	juggl	juggle+d	_jugg, juggl, uggle, ggled, gled_
jugglers	juggler	juggle+r+s	_jugg, juggl, uggle, ggler, glers, lers_

All Methods

	words	stems	morf	lcn4	lcn5	4-grams	5-grams
BG	0.2164		0.2703	0.2822	0.2442	0.3105	0.2820
CS	0.2270		0.3215	0.2567	0.2477	0.3294	0.3223
DE	0.3303	0.3695	0.3994	0.3464	0.3522	0.4098	0.4201
EN	0.4060	0.4373	0.4018	0.4176	0.4175	0.3990	0.4152
ES	0.4396	0.4846	0.4451	0.4485	0.4517	0.4597	0.4609
FI	0.3406	0.4296	0.4018	0.3995	0.4033	0.4989	0.5078
FR	0.3638	0.4019	0.3680	0.3882	0.3834	0.3844	0.3930
HU	0.1520		0.2327	0.2274	0.2215	0.3192	0.3061
IT	0.3749	0.4178	0.3474	0.3741	0.3673	0.3738	0.3997
NL	0.3813	0.4003	0.4053	0.3836	0.3846	0.4219	0.4243
PT	0.3162		0.3287	0.3418	0.3347	0.3358	0.3524
RU	0.2671		0.3307	0.2875	0.3053	0.3406	0.3330
SV	0.3387	0.3756	0.3738	0.3638	0.3467	0.4236	0.4271
Average	0.3195		0.3559	0.3475	0.3431	0.3851	0.3880
	0.3719	0.4146	0.3928	0.3902	0.3883	0.4214	0.4310

TEL Experiments (Monolingual)

- **Treated records as 'free text'**
  - **Removed some metadata fields (e.g., publisher, rights, spatial)**
  - **Removed tag structure**

	English	French	German	Run Description
words	0.2719	0.2019	0.1073	
stems	0.3480	0.2290	0.1757	aplmoxxs
morf	0.3171	0.2332	0.1989	
lcn4	0.3086	0.2223	0.1565	
lcn5	0.2993	0.2270	0.1810	
4-grams	0.3382	0.2950	0.3377	aplmoxx4
5-grams	0.3190	0.2800	0.3102	aplmoxx5
4-grams+RF	0.3531	0.2861	0.3176	aplmoxx4rf

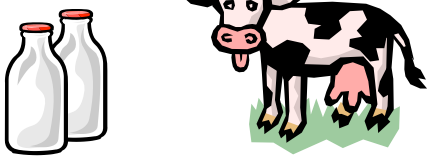
human language technology
center of excellence

Subword Translation

- Like words, n-grams can be translated using aligned text
- Attempt is to mitigate problems in dictionary-based CLIR
  - word lemmatization (variation in morphology)
  - out of vocabulary words, particularly names (no OOV n-grams)
  - multiword expressions (due to word-spanning n-grams)

	German	Italian
word	milch	latte
stem	milch	latt
4-grams	milc ilch	latt latt
5-grams	_milc milch ilch_	_latt _latt latte

	French	Dutch
word	lait	melk
stem	lait	melk
4-grams	lait	melk
5-grams	_lait lait_	_melk melk_



TEL Experiments (Bilingual)

- 5-grams in query 'translated' from source to target
  - No pre-translation query expansion
  - No automated relevance feedback
- Parallel corpora: OJEU 1998-2004

	English	French	German
Dutch	0.2024	0.1746	
English	x	0.1609	0.1899
French	0.2087	x	0.1829
German	0.2111	0.1608	x
Spanish	0.1856		

65.4% 62.4% 61.2%

Character Skipgrams

- Character n-grams: robust matching technique
- Skipgrams: super robust matching
  - Some letters are omitted (essentially a wildcard match)
- Skip bi-grams for fuzzy matching
  - Pirkola et al. (2002): learning cross-lingual translation mappings in related languages
  - Mustafa (2004): monolingual Arabic retrieval
- This work generalizes approach to consider longer n, multiple (non-adjacent) skips
  - sw*m matches swim / swam / swum
  - f**t matches foot / feet
- Application to OCR'd docs and complex morphology

Skipgram Indexing

- Example: 4,2 skipgrams for Hopkins
  - 4 letters, 2 skips
  - hkin, hpin, hpkn, hoin, hokn, hopn
  - oins, okns, okis, opns, opis, opks
  - Note: more skipgrams than plain n-grams
- Initial work uses combined index
  - Plain 4-grams (e.g., 'hopk')
  - 4,1 skipgrams (e.g., 'h*pk')
- Preliminary results show small improvement in Hungarian and Czech
- What about in Persian?

Persian Results

- **Official runs: words, morf, 5-grams, sk41**
- **Bilingual runs based on MT**
  - **http://www.parstranslator.net/eng/translate.htm**

	Task	RF terms	MAP	Run ID
5-grams	mono	100	0.4493	jhufa5r100
sk41	mono	400	0.4519	jhufask41r400
words	mono	50	0.4332	jhufawr50
morf	mono	50	0.4250	jhufamr50
5-grams	bi	100	0.1660	jhuenfa5r100
sk41	bi	400	0.1892	jhuenfask41r400
words	bi	50	0.0946	jhuenfawr50
morf	bi	50	0.1112	jhuenfamr50

Summary

- Studied diverse tokenization methods using TEL and in all CLEF ad hoc news test sets
- Regular N-grams are the single best solution
  - 21% gain over words
  - Improvement greatest in morphologically richer languages
  - Key advantage appears to be morphological normalization
- Other methods have merit
  - Stemming works well, if available, particularly in Romance languages
  - Morfessor segments and n-gram 'stems' are efficient alternatives to n-grams when no stemmer is available
- Skip n-grams interesting future direction