USING SYLLABLES AS INDEXING TERMS IN FULL-TEXT INFORMATION RETRIEVAL

Kimmo Kettunen, Paul McNamee and Feza Baskaya

HLT2010, Riga, October 7-8, 2010

Outline

1. Why use syllables?
2. Our view of syllabification
3. IR test collections
4. Results
5. Discussion & conclusions

Why use syllables?

- N-gramming has been found very effective in handling of different languages in IR (e.g. P. McNamee and J. Mayfield, Character n-gram tokenization for European language text retrieval, *Information Retrieval* **7** (2004), 73–97.) N = 2-6 chars
- Syllables resemble n-grams, but there are less of them and their length varies
- Syllables have been used much in speech retrieval but not much in text retrieval
- There are syllabifiers around, and it is also quite simple to write a simplified syllabifier for a language
- Perhaps one simplified syllabifier works for different languages even?

Our view of syllabification

- Syllabification as a linguistic problem is trickier than thought, because views of syllable structure vary; thus there might be different syllabifications for words in different languages

- Algorithmic syllabification can be rule-based or data-driven; nowadays data-driven methods are popular and seem also to be efficient. Typical accurary rates for syllabification are over 95 %, best over 99 %

- N. B. there does not seem to be gold standard collections for syllabification of different languages, so evaluation of syllabification algorithms is not on the same level as e.g. evaluation of morphological processing

Our view of syllabification

- Most of the languages have one basic syllable structure: CV, consonant + vowel
- We had two basic syllabification strategies:
  - 1) put a hyphen after every CV
  - 2) put a hyphen before every CV
- CV_1 (*ca + rbo + hy + dra + te + s; do + gs; go + es)*
- *CV_2 (car+bo+hyd+ra+tes; dogs; goes)*
- *These two procedures were tried with 14 languages*
- *With 3 languages we tried also proper syllabification programs*

IR test collections

- Cross-language Evaluation Forum (CLEF) data for 13 languages (**BG, CS, DE, EN, ES, FI, FR, HU, IT, NL, PT, RU, SV)** + Milliyet collection for Turkish
- The size of the CLEF collections vary from ~17 000 to 450 000 documents. The number of topics for each collection is between 50 and 367; Milliyet has 408 305 documents and 72 topics
- Title + description queries (= long queries) were run for all the languages
- Retrieval engines: HAIRCUT for CLEF, Lemur for Milliyet
- Baseline: plain words; comparable methods: Snowball stemming, 4-gramming

Results with simple syllabification, 14 languages

Table 1. Results of CV_1 and CV_2 syllabi[illegible] runs for 14 languages, title and description queries, mean average precisions (MAPs)

	words	snow	4	syl1_CV1	syl2_CV1	syl3_CV1	syl1_CV2	syl2_CV2	syl3_CV2
BG	0.22	N/A	**0.31**	0.21	0.22	0.10	0.21	0.20	0.10
CS	0.23	N/A	**0.33**	0.18	0.26	0.16	0.19	0.27	0.19
DE	0.33	0.37	**0.41**	0.28	0.39 ←	0.29	0.30	0.38 ←	0.24
EN	0.41	0.44	**0.40**	0.21	0.38	0.27	0.23	0.35	0.20
ES	0.44	0.49	**0.46**	0.24	0.45	0.31	0.22	0.43	0.29
FI	0.34	0.43	**0.50**	0.30	0.46 ←	0.38	0.27	0.43 ←	0.31
FR	0.36	**0.40**	0.38	0.20	0.37	0.25	0.23	0.34	0.22
HU	0.20	N/A	**0.38**	0.20	0.32 ←	0.23	0.18	0.29 ←	0.18
IT	0.38	**0.42**	0.37	0.18	0.39	0.26	0.17	0.37	0.26
NL	0.38	0.40	**0.42**	0.26	0.38	0.25	0.29	0.36	0.23
PT	0.32	N/A	**0.34**	0.17	0.33	0.20	0.17	0.30	0.16
RU	0.27	N/A	**0.34**	0.28	0.24	0.13	0.26	0.26	0.15
SV	0.34	0.38	**0.42**	0.26	0.41 ←	0.31	0.25	0.37 ←	0.26
TU	0.19	0.22	**0.31**	0.17	0.30 ←	0.22	0.21	0.26 ←	0.20

Table legend: words = surface forms (lower-cased); snow = Snowball stemmer; 4 = overlapping, word-spanning character 4-grams; syl1 = single syllables; syl2 = syllable bigrams; syl3 = syllable trigrams.

Results

Table 2. Averages and changes from plain words baseline

	words	snow	4	syl1_CV1	syl2_CV1	syl_CV1	syl1_CV2	syl2_CV2	syl3_CV2
Avg-8	0.37	0.42	0.42	0.24	0.40	0.29	0.25	0.38	0.25
Chg-8 %	N/A	11.47	13.31	-34.69	7.89	-22.18	-34.14	1.60	-32.80
Avg-A	0.32	N/A	0.39	0.23	0.35	0.24	0.23	0.33	0.21
Chg-A %	N/A	N/A	20.54	-29.15	8.69	-25.25	-29.42	3.40	-33.75

Table legend: Avg-8 is average over 8 'Snowball' languages, i.e. languages that had available Snowball stemmer; Avg-A is average over the CLEF data; Chg-A is change over plain words with the CLEF data average.

Results

- For three languages we had proper syllabification algorithms: De, Fi, Tu

	Syl1	Syl2	Syl3
De	0.31	0.36	0.33
Fi	0.28	0.44	0.33
Tu	0.21	0.27	0.20

Results

- Statistically significant relative gains vs. surface forms in four languages using syllable bigrams with CV_1 procedure:

- German (+18.5%, relative)
- Finnish (+34.8%)
- Hungarian (+60.4%)
- Swedish (+19.9%).

- With Turkish the CV_1 procedure with *syl2* was performing at the same level as 4-grams, which is interesting.

Proper syllabification did not outperform CV_1, but performed relatively well with syllable bigrams

Results

Sizes of indexes, examples

		DE	EN	FI	SV
	CLEF year	2003	2007	2004	2003
	Docs	294805	87653	55344	142819
words	total	85057494	49956344	14394166	29218580
	unique	1180570	214742	975390	498858
	avglen(type)	12,63	7,50	12,61	11,24
	avglen(token)	5,98	4,68	7,23	5,26
4-grams	total →	583763758	279021872	116051783	179480821
	unique	216918	149010	136224	160299
	avglen(type)	4,00	4,00	4,00	4,00
	avglen(token)	4,00	4,00	4,00	4,00
CV1	total →	216793161	103899845	43804039	68411444
	unique	80948	37180	26679	48701
	avglen(type)	4,71	4,30	4,10	4,40
	avglen(token)	2,35	2,26	2,38	2,25
	syl-per-word	2,55	2,08	3,04	2,34
CV2	total	175651493	92536770	39814528	59393259
	unique	78462	40604	30396	48755
	avglen(type)	4,68	4,34	4,16	4,38
	avglen(token)	2,90	2,53	2,62	2,59
	syl-per-word	2,07	1,85	2,77	2,03

Notes
1. Total = number of term occurrances in document collection
2. Unique = number of unique terms in collection
3. Average lengths are in characters; by type (lexicon) and token (collection).
4. syllables-per-word are computed by dividing total syllables by total words

Discussion & conclusions

- Overall our results show that syllables can be used effectively in management of word form variation for different languages. They are not able to outperform 4-grams, but at best they perform at the same level or slightly better than Snowball stemmer for morphologically complex languages, such as Finnish, German, Hungarian, Swedish and Turkish. → **This is a good result**

- As with n-grams, there seems to be a an optimal length for items put in the index : bigram syllables. These result on index items of 4-5 characters on average. These items do take care of morphological variation relatively well

- A simple CV procedure does not suit all the languages: it is not language independent, but at least it is flexible with languages.

Remember this!

- One simplified syllable algorithm handled 5 morphologically complex languages well IR wise!
- It suits also morphologically easier languages, but there is not as much to be gained