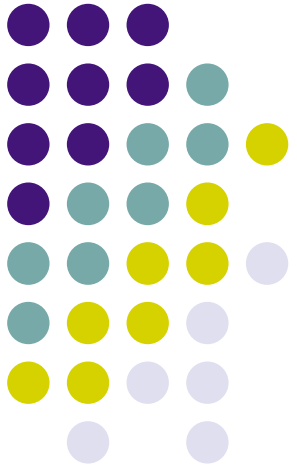
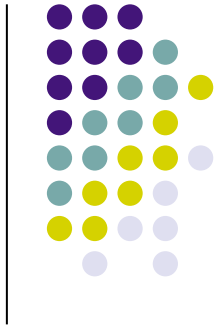
Managing Word Form Variation of Text Retrieval in Practice – why Five Character Truncation Takes it all?

Kimmo Kettunen

Human Language Technologies – The Baltic Perspective, Tartu,Oct 4-5, 2012

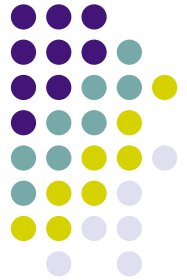


Background



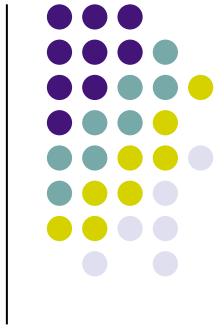
- Natural language is a problem in textual information retrieval
- There are several layers of problems, but we concentrate on morphology (i.e. variation of word forms, especially nominals)
- Discussion about different methods for word form variation management, practical orientation (IR goals vs. NLP goals)

Word form variation



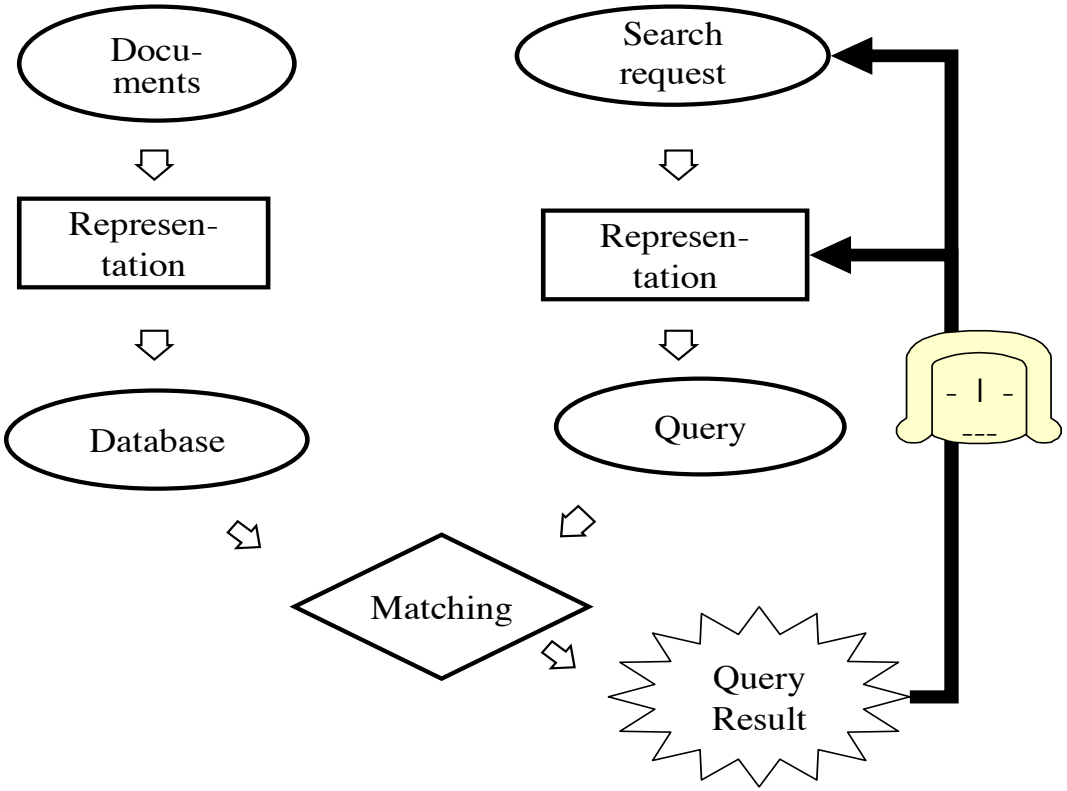
- Inflectional morphology is one of the main reasons for word form variation
- Complexity of inflectional systems in languages differs
- *Roughly*: English ←---→ Finnish (Estonian), and stops in the between (and beyond)
- E.g. number of cases: English 2 – Finnish/ Estonian 14: a totally different game

IR basics 1

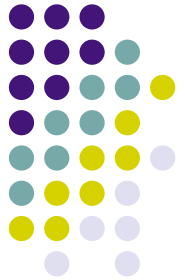


- Full-text information retrieval aims in retrieving relevant documents for the user ranking the most relevant at top of the list
- Relevance / aboutness (a very fuzzy concept)
- In practice user gives keywords, i.e. search terms that somehow describe his/her need for information, for the search engine
- Search engine works on the basis of this given input

IR basics 2

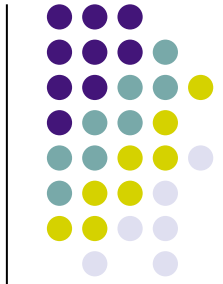

Docu-
ments
Represen-
tation
Database
Search
request
Represen-
tation
Query
Matching
Query
Result

IR basics 3



- Matching of the input keywords (representations) to the database (index) keywords (representations)
- Matching is a string comparison process based on similarity (no semantics, just strings)
- *Cat – cats* → no match
- *Poika, pojan, pojassa…* → no match
- *Poeg, poja, poega…* → no match

Management of word form variation



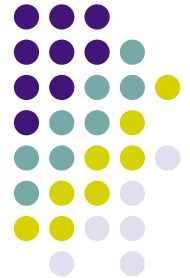
- Variation needs to be taken care of for IR engines
- Lots of methods have been developed to take care of variation (i.e. make representations similar)
- **E.g**.: term truncation, stemming, character n-gramming, lemmatization, inflected form generation etc.
- Some of these are rough, some more linguistic methods
- Differences in IR performance **not that big** many times

McNamee, Nicholas, Mayfield 2009

Table 1: Examples of indexing term formation.

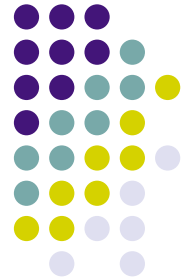
words	medical, doctors
snow	medic, doctor
morf	medical, doctor, s
devowel	m.d.c.l, d.ct.rs
soundex	5030204, 3023062
lfs4	edic, doct
lfs5	medic, docto
trun4	medi, doct
trun5	medic, docto
3-grams	_me, med, edi, dic, ica, cal, al_, l_d, _do, ...
4-grams	_med, medi, edic, dica, ical, cal_, al_d, l_do, _doc, doct, ...
5-grams	_medi, medic, edica, dical, ical_, cal_d, al_do, l_doc, _doct, ...
6-grams	_medic, medica, edical, dical_, ical_d, cal_do, al_doc, l_doct, ...
7-grams	_medica, medical, edical_, dical_d, ical_do, cal_doc, al_doct, ...
sk41	regular 4-grams in addition to m.dic, me.ic, med.c, e.ica, ed.ca, edi.a, ..., a._do, al.do, al_.o, l.doc, l_.oc, l_d.c, ...
wisk41	word-internal 4-grams in addition to m.dic, me.ic, med.c, e.ica, ed.ca, edi.a, ...,
win4	_med, medi, edic, dica, ical, cal_, _doc, doct, octo, ctor, tors, ors_
win5	_medi, medic, edica, dical, ical_, _doct, docto, octor, ctors, tors_

McNamee, Nicholas, Mayfield 2009 - Results shortly



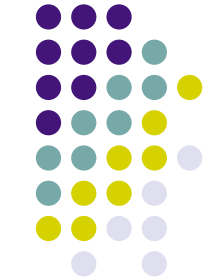
- 18 different methods for 18 languages evaluated in 5 different writing systems
- character n-gramming is **the most effective** method for most of the languages
- rule based stemming (Snowball stemmers are used) can be an attractive option for languages where morphological variation is not very high
- phonetic transformations do not work well for any language

McNamee, Nicholas, Mayfield 2009 - Results shortly



- a statistical stemmer (i.e. particular unsupervised morphological method) does not perform too well, but is getting better (cf. also Kurimo et al. [5] for the latest results with different systems)
- one of the most unsophisticated and un-linguistic methods, **five character truncation**, works very well with most of the languages, being the second best non n-gram method overall, only slightly behind performance of Snowball stemmers.

How to measure success of IR?



- IR success is measured with performance measures: mostly effectiveness of search
- These measures quantify, how many relevant and irrelevant documents are retrieved
- AP = average precision
- MAP = mean average precision
- and other measures…(GMAP, P(10), DCG…)

What do different languages gain from morphological processing?

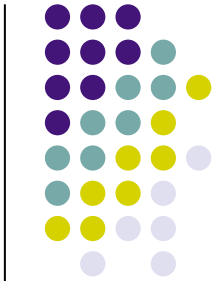
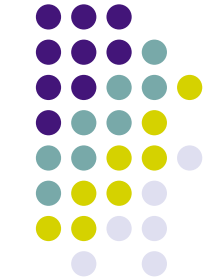


Table 1. Necessity of word form variation management in the light of MAP results

Language	GAP = best MAP with word form variation management *minus* plain words MAP [6]	Lowest and highest MAPs gained [10]		Is word form variation management needed for the language?
		low	*high*	
1. Bulgarian	6.8-8.1 %	0.216	0.31	beneficial
2. Czech	N/A	0.227	0.329	necessary
3. Dutch	0.6.-5.0 %	0.381	0.424	no need
4. English	1.2-2.9 %	0.406	0.437	no need
5. Finnish	10.5-25.2 %	0.34	0.507	necessary
6. French	0.5-3.8 %	0.363	0.401	no need
7. German	6-15.7 %	0.33	0.42	beneficial/necessary
8. Hungarian	9.9-12.4 %	0.197	0.374	necessary
9. Italian	N/A	0.374	0.417	no need
10. Portuguese	N/A	0.316	0.352	no need
11. Russian	6.1-21.0 %	0.267	0.373	necessary
12. Spanish	N/A	0.439	0.484	no need/beneficiary
13. Swedish	1.7-8.8 %	0.338	0.427	beneficial
14. Turkish	12.3 %	N/A[1]		necessary

Criteria for choosing word form variation management method



How to choose the method for word form variation management used with text IR?

1) Take a look at the complexity of the language you need to handle in the IR engine

E.g. there must be hundreds of papers concerning English IR from the viewpoint of word form variation management? Why? The language is morphologically simple, not much can be gained anyhow whatever you do.

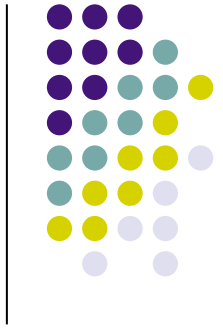
Why bother at all for a few per cent gain?

Criteria for choosing word form variation management method

Table 2. Scoring of different word form variation management methods along five criteria

Method	Language independence	Effective-ness	Index size	Automatic generation of rules	Simplicity of the approach	SUM
automatic truncation [10]	4	3.33	3.33	2.66	4	**17.32**
unsupervised morphological methods [4,5]	4	4	3	2	2	**15.0**
syllabification [11]	3.33	2.66	2	2.66	3.33	**13.98**
n-gramming (plain, no skips) [10]	4	3.33	0	2	3.33	**12.66**
statistical lemmatization [19]	2.66	3.33	1.33	2	1.33	**10.65**
rule based stemming [10]	0.66	2.66	3.33	0.66	2	**9.31**
plain words	4	0	1.33	0	3.33	**8.66**
word form generation [6]	0.66	4	1.33	1.33	1.33	**8.65**
lemmatization (rules + dict.) [3]	0	4	2	0	0.66	**6.66**

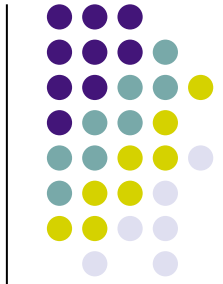
Heuristics: a suggestion



1) For morphologically simple languages (such as 3, 4, 6, 9 in Table 1) do nothing but normal routines (case folding etc.). Plain word forms are a good solution for indexing and query formation with these languages.

2) If the language is in the *beneficial* group (such as 1, 7, and 13 in Table 1), the simplest non-linguistic word form management method can be used. Out of the simple methods five character truncation is the easiest to implement and very effective, but also n-gramming and hyphenation could be used. Large indexes and slow retrieval are shortcomings of n-gramming. A light stemmer can also be considered, if such is available. But there is no need for ‘heavy artillery’ here.

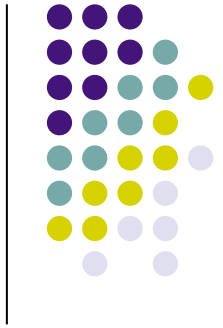
Heuristics: a suggestion



3) With languages in the *necessary* group *(*such as 2, 5, 8, 11 and 14 in Table 1) one can begin to consider also ‘heavier’ methods, such as stemming or lemmatization. Even here they are not necessary, as five character truncation is effective with these languages too.

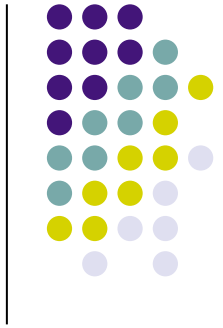
If one’s only need is to have the best IR performance from the search engine, then language technology oriented tools **may be overkill**. If one has also other needs for the linguistic analysis capabilities of the IR system in whole (such as handling of lemmas or interaction as e.g. in query expansion, cf. Galvez et al. [15], then one may consider an elaborate lemmatizer.

Conclusion

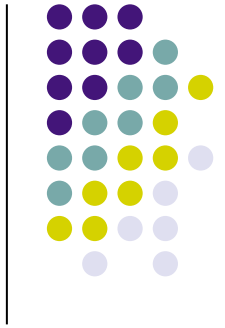


- Present IR engines are not very clever. Anyhow they give very good search results (cf. Google, Bing)
- NLP methods aim at linguistic completeness
- Text IR is a fuzzy process, where also very rude word form handling works well many times
- You need to think about what you really need when choosing your word form variation management method
- Remember: you are not doing mainly NLP in IR! NLP is a helper, not an aim in itself.

Conclusion



- **MDL** = Minimal Description Length
- when two models fit the data equally well, MDL will choose the one that is the simplest in the sense that it allows for a shorter description of the data
- E.g. five character truncation could be favored instead of a lemmatizer, as it is far simpler and “fits the data” – i.e. management of word form variation for IR – as well as stemming or lemmatization with many languages.



Thank you very much. Any questions?