LIGHT STEMMING FOR ARABIC INFORMATION RETRIEVAL

Nasser Alansari
SWE

Outlines

1. Introduction
2. Review of 2002 Stemming experiments
3. New Studies of stemming Via Morphological Analysis
4. Conclusions

Introduction

1. Arabic Morphology & Orthography
2. Stemming in IR
3. Stemming & Morphological Analysis in Arabic

Arabic Morphology & Orthography

- Morphology complexity of Arabic:
  - Difficulty to develop NLP application for Arabic IR
  - Most noun, adjective and verb stems derived from thousand roots by infixing:
    - E.g. maktab , kitAb , kutub , kataba , naktubu => from ktb(root)

Arabic Morphology & Orthography (cont)

- Arabic orthography:
  - can confuse IR system
  - Arabic cab be written with or without the diacritics:
    - E.g. كَتَبَ and كتب look similar to eye, but to computer they don't match
  - Orthography with diacritics is less ambiguous and more phonetic
  - Diacritics text can only found in specialized contexts:
    - E.g. The Qur'an, Children's book , Dictionaries

Stemming in IR

- **Stemming**: is tool that used in IR to combat vocabulary mismatch problem.
- Two classes for stemming approaches:
  - Affix Removal
  - Statistical Stemming (e.g n-grams)

Stemming & Morphological Analysis in Arabic

- The factors introduced in “Arabic morphological & Orthography” make Arabic very difficult to stem
- Approaches for stemming in Arabic:
  - Manual Construction of Dictionaries
  - Affix Removal (light stemming)
  - Statistical Stemming
  - Morphological Analysis

Stemming & Morphological Analysis in Arabic (cont)

- Manual Construction of Dictionaries
  - Early approach
  - Al-Kharashi and Evens worked with small text collections
  - They manually built dictionaries of roots and stem for each word to be indexed
  - This approach is obviously impractical for realistic sized corpora.

Stemming & Morphological Analysis in Arabic (cont)

- Affix Removal
  - It generally called *light stemming*
  - It a process of stripping off a small set of prefixes and/or suffixes
  - Without dealing with infixes or recognize patterns and finds roots
  - Light10 is light stemmer approach

Stemming & Morphological Analysis in Arabic (cont)

- Light10:
  - Is one of the light stemmer approach
  - Strips off initial ‘و’
  - And definite articles (ال ، وال ، بال ، كال، فال، لل)
  - And suffixes (ي ، ة ، ه ، ية ، يه ، ين ، ون ، ان ، ها)
  - It was designed to strip off strings that were frequently found as prefixes or suffixes.
- Al-Stem:
  - Darwish introduced it in TREC (2002)
  - Less effective that light10
- Chan & Grey
  - Introduced a light stemmer similar to light10
  - Remove more prefixes and suffixes
  - More effective than Al-Stem

Review of 2002 Stemming experiments

1. Experimental Method
2. Light Stemmers
3. Results of Morphological Stemmer Comparisons
4. Comparison With Morphological analysis
5. Cross-language Retrieval

Experimental Method

- TREC-2001 Arabic corpus
  - Also called the AFP_ARB
  - Consists of 383,872 newspaper articles in Arabic
  - From Agency France Press
  - Fill up almost gigabyte in UTF-8 encoding
- Corpus and queries
  - Converted to CP1256 encoding
  - Indexed using in-house version of INQUERY retrieval engine

Experimental Method

- Raw condition
  - Where no normalize and stemming used
  - Corpus and queries are normalized according to:
    - Remove non letters
    - Replace إ،أ،آ with ا
    - Replace final ى with ي
    - Replace final ة with ه

Light Stemmers

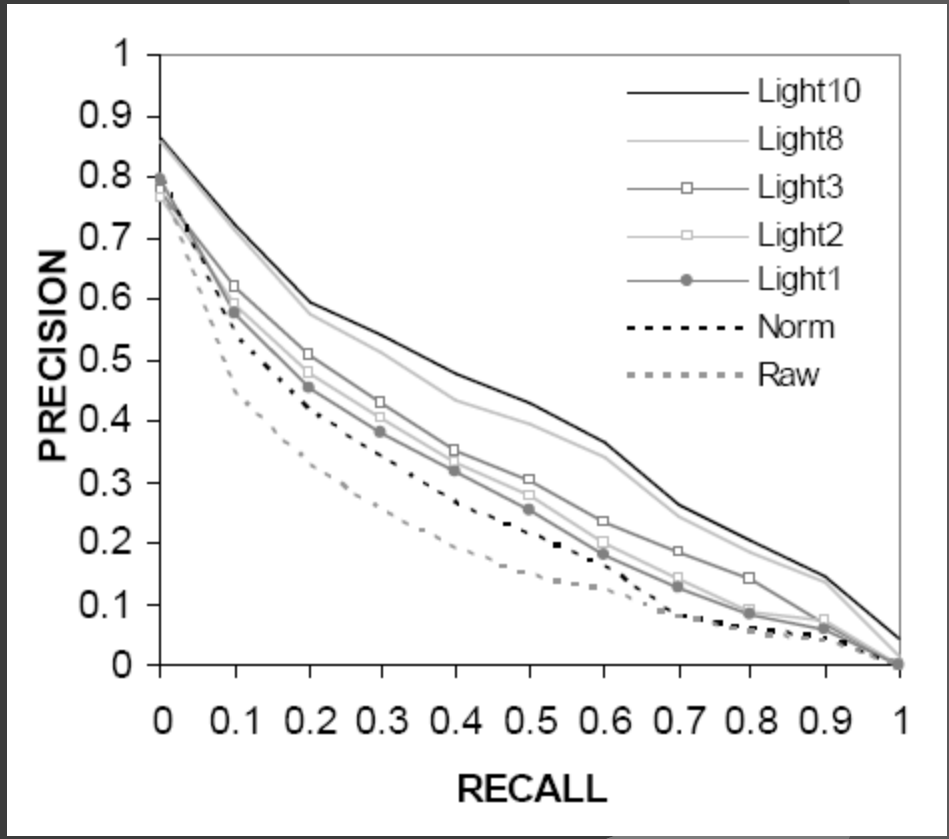
- Steps to be apply to all version of light stemmers:
  1. Remove و from lgiht2, light3, and light8
     1. And light10 if the remainder of the word is 3 or more characters long
  2. Remove any definite article if this leaves 2 or more characters
  3. Go through the list of suffix once in RTL order (Table 1).

Light Stemmers

	Remove prefixes	Remove Suffixes
Light1	ال، وال، بال، كال، فال	none
Light2	ال، وال، بال، كال، فال، و	none
Light3	"	ه، ة
Light8	"	ها، ان، ات، ون، ين، يه، ية، ه، ة، ي
Light10	ال، وال، بال، كال، فال، لل، و	"

Results of Morphological Stemmer Comparisons

- **Raw**: mean no normalization or stemming.
- **Norm**: mean normalization but no stemming.
- **LightX**: refer to light steamers



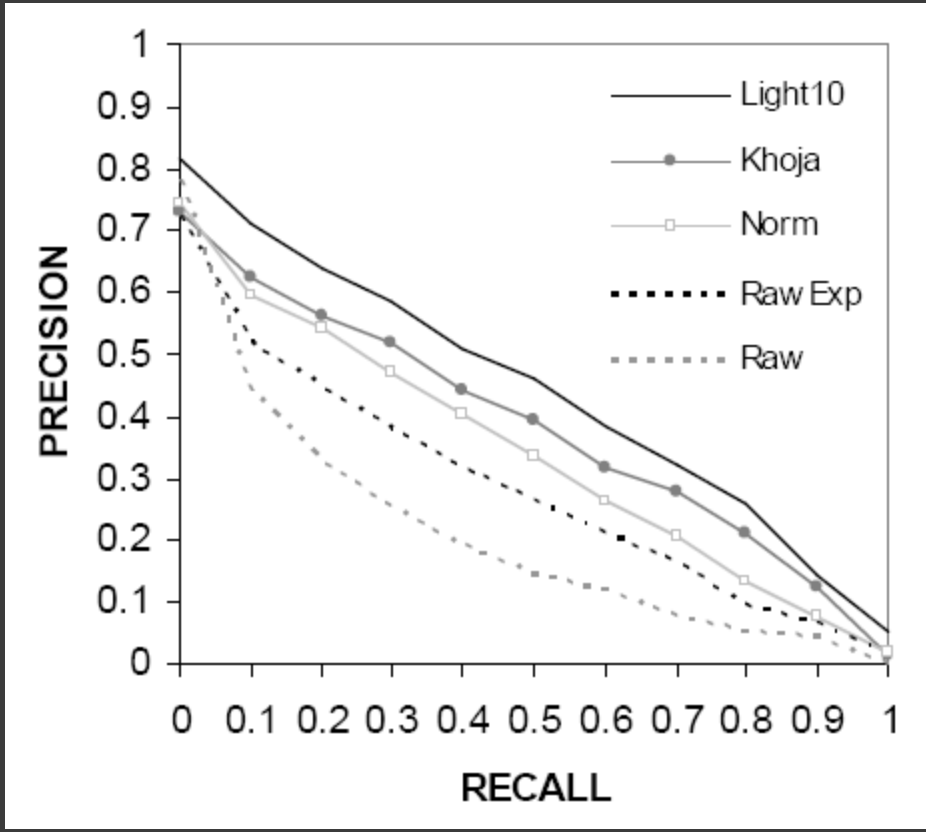
Results of Morphological Stemmer Comparisons (cont)

Stemmer	raw	norm	light1	light2	light3	light8	light10
Average Precision	.196	.241	.273	.291	.317	.390	.413
Percent Change		22.9	39.3	48.3	61.8	98.7	100.1

Comparison With Morphological analysis

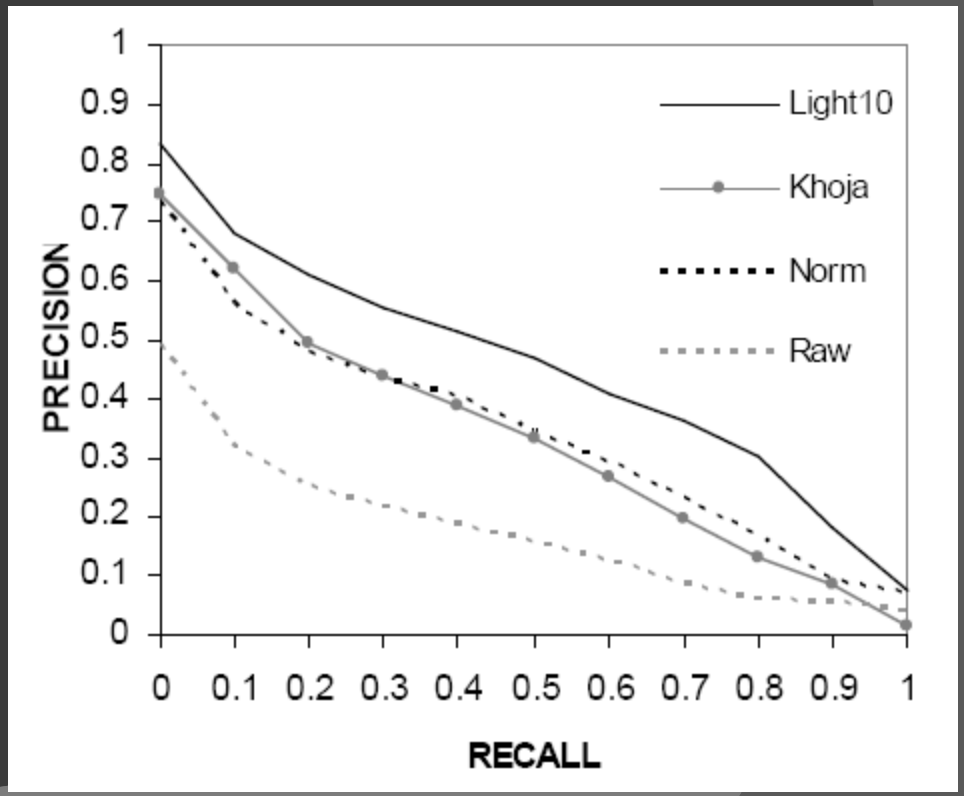
- Khoja stemmer was used to find roots for indexing and retrieval
- The average precision for Khoja stemmer is .341
  - Significantly worse then light10

Comparison With Morphological analysis (cont)



Cross-language Retrieval

- Khoja morphological analyzer and Loght10 also compared in Cross-language Retrieval



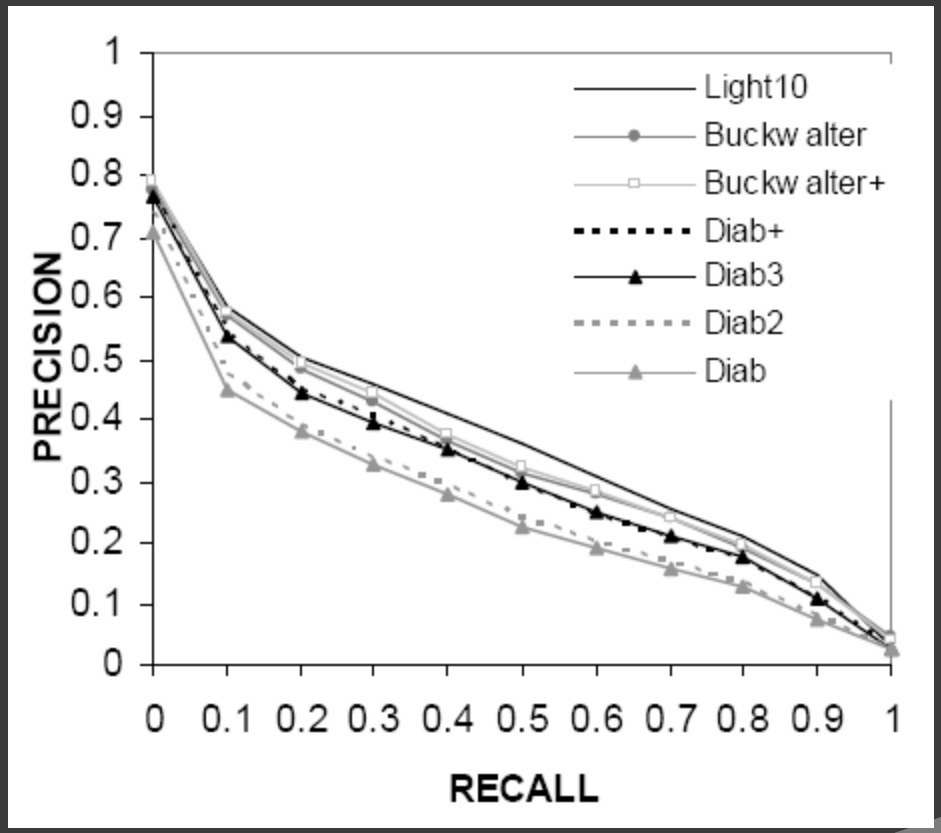
Cross-language Retrieval

Stemmer	raw	norm	khoja	light10
Average Precision	.113	.262	.260	.384
Percent Change		133	130	240
With English Query Expansion				
Average Precision	.139	.306	.308	.425
Percent Change		120	121	206
With English and Arabic Query Expansion				
Average Precision	.163	.336	.321	.447
Percent Change		106	97	174

New Studies of stemming Via Morphological Analysis

1. Buckwalter Morphological Analyzer
2. Diab Tokenizer, Lemmatizer and POS Tagger
3. Comparison with Light Stemmer

Comparison with Light Stemmer



Comparison with Light Stemmer

Stemmer	Light10	Buckwalter	Buckwalter+	Diab	Diab2	Diab3	Diab+
Unexpanded	.353	**.330**	**.334**	**.247**	**.257**	**.302**	**.302**
Expanded	.387	.386	.390	**.322**	**.336**	**.354**	**.356**

Conclusions

- Stemming has a large effect on Arabic IR
  - Far large than the effect in other language
- The stemmer was a light stemmer (light10)
- Why would a morphological analyzer not perform better than such a simple stemmer?
  1. Morphological analyzer make mistakes (particularly on names).
  2. Models used in IR treat document and queries as “bags of word” or “bags of unigrams, bigrams, trigrams”
  3. Light stemming is robust (not requires complete sentences)
  4. It is still not clear what the correct level of conflation should be for IR

Thanks

1 + 1 = ?!