Approaches to Arabic Information Retrieval

Jihad Ashkar

Outline

- Challenges in Arabic IR
  - Vocalization
  - Derivation and Inflection
  - Irregular Nouns
- Approaches to Arabic IR
  - Normalization
  - Stemming
  - Morphology Analysis
  - N-Grams
- Comparison of Approaches
- Future Work

Challenges in ARABIC IR

- Vocalization
  - Use of diacritic marks can give words different meanings

عِلْم	Ambiguous
عَلَم	Flag
عِلِم	Science
عَلَّمَ	Taught

Challenges in ARABIC IR

- Derivational and Inflectional Nature of Arabic
  - Many definite articles, conjunctions and other prefixes can attach to the beginning of a word, and large numbers of suffixes can attach to the end.
  - New words can be formed by transforming root words
  - According to Hegazi and Elsharkawi (1985) there are about 1200 roots.
  - It is possible to derive 120 different forms of nouns by adding affixes to the basic naked noun, and 1440 different forms of verbs by adding affixes to the basic naked verb.

Challenges in Arabic IR

Arabic word	English translation	Arabic word	English translation	Arabic word	English translation	Arabic word	English translation
أطفال	children	اطفالهن	their children	بطفل	by child	فالطفلة	then the child
أطفالا	children	اطفالي	my children	بطفلة	by child	فطفل	then child
أطفالنا	our children	الأطفال	children	بطفلتنا	by our child	كأطفال	as children
أطفاله	and his children	الاطفال	children	بطفلته	by his child	كالطفل	as the child
أطفاله	his children	الطفل	the child	بطفله	by his child	لأطفال	to children
أطفالها	her children	الطفلان	the children	بطفلها	by her child	لطفلها	to her child
أطفالهم	their children	الطفلة	the child	بطفلهما	by their child	للطفلة	to the child
أطفالهن	their children	الطفلتان	the children	بطفلين	by children	وأطفالنا	and our children
أطفالي	my children	الطفلتين	the children	بطفليها	by her children	والأطفال	and the children
اطفال	children	الطفله	the child	طفل	child	وبطفل	and by child
اطفالا	children	الطفلين	the children	طفلا	child	وبطفلين	and by children
اطفالك	your children	بأطفال	by children	طفلان	children	وطفلة	and child
اطفالكم	your children	بأطفاله	by his children	طفلاها	her children	وطفلتان	and children
اطفالكن	your children	بأطفالها	by her children	طفلة	child	وطفلنا	and our child
اطفالنا	our children	بالأطفال	by the children	طفلت	child	وطفلها	and her child
اطفاله	his children	بالطفل	by the child	طفلتان	children	وطفليه	and his children
اطفالها	her children	بالطفلة	by the child	طفلتة	his child	وطفليها	and her children
اطفالهم	their children	بالطفلتين	by the children	طفلتنا	our child	ولأطفالها	and to her children
اطفالهما	their children	بالطفلين	by the children	طفلته	his child	وللطفل	and to the child

Table 1: Arabic words whose English translations contain the headword *child* or *children*.

Challenges in ARABIC IR

- Irregular Single and Plural Nouns

Regular Nouns	
معلّم (teacher, masculine)	معلّمين (teachers, masculine)
مدرّبة (trainer, feminine)	مدرّبات (trainers, feminine)

Irregular Nouns	
طفل (child)	أطفال (children)
امرأة (woman)	نساء (women)

Approaches to Arabic IR

- Normalization
  - Convert to encoding of choice
  - Remove punctuation
  - Remove diacritics (weak vowels)
  - Remove non-letters
  - Change letters to normalized form

Approaches to Arabic IR

- Morphological Analysis
  - Khoja and Garside (1999)
  - Removes prefixes and suffixes
  - Attempts to map word to root

Approaches to Arabic IR

- Stemming
  - Light stemming
    - Remove subset of prefixes or suffixes
      - Predefined or determined from corpus
  - Simple stemming
    - Remove vowels from words
  - MT-based Stemmer
    - Clustering based on English translations

	Remove from front	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، يه، ه، ة، ي

Approaches to Arabic IR

- Character n-grams
  - 3 characters, 4 characters, 5 characters
  - Beginning of word, end of word, middle of word
  - With crossing, without crossing

Comparison of Approaches

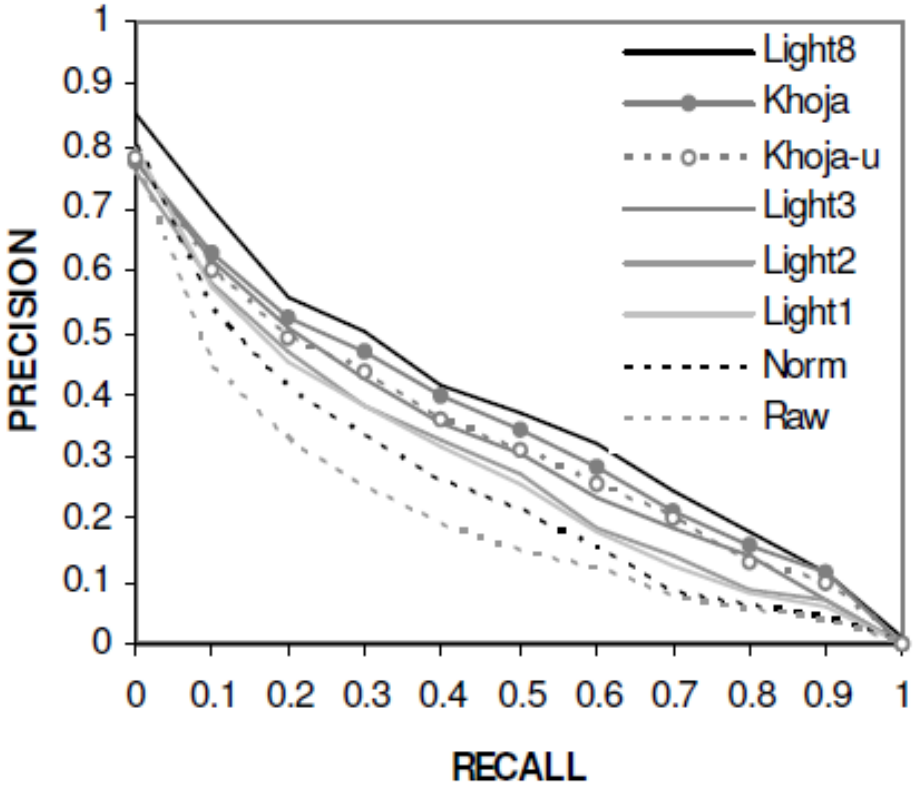

Figure 1: Monolingual 11 point precision for basic stemmers, unexpanded queries

Comparison of Approaches

Table 2: Monolingual average precision for basic stemmers, unexpanded

Stemmer	raw	norm	light1	light2	light3
Av. Precision	.194	.238	.273	.284	.317
Pct. Change		23.1	41.1	46.7	63.9

Stemmer	raw	khoja-u	khoja	light8
Av. Precision	.194	.313	.341	.376
Pct. Change		61.7	76.2	94.3

Comparison of Approaches

			without expansion		with expansion	
run id	stemmer	index unit	recall	precision	recall	precision
mon0	NONE	word	4035	0.2365	4583	0.2872
mon1	NONE	trigram (without crossing)	3914	0.2398	4632	0.3239
mon2	NONE	trigram (with crossing)	4018	0.2479	4681	0.3178
mon3	Al-Stem stemmer	word	4500	0.2858	4864	0.3482
mon4	MT-based stemmer	word	4402	0.2948	4885	0.3348
BKYMON	Berkeley light stemmer	word	4543	0.3099	4952	0.3666

Table 4: Monolingual retrieval performances. The number of relevant documents for all 50 topics is 5909. Only the *title* and *description* fields were indexed.

Comparison of Approaches

a) N-gram (3 characters)

Dice's coefficient (threshold =0.6)			Cosine coefficient (threshold =0.7)			TF*IDF weight (threshold=0.01)		
Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
53.3	40.81	46.22	41.34	10.8	21.68	37.34	26.9	31.27
92.85	1.45	2.85	89.28	2.34	4.56	46.42	37.7	41.61
73.075	21.13	24.53	65.31	6.57	13.12	41.88	32.3	36.44

Table 1: result using 3-gram

b) N-gram (4 characters)

Dice's coefficient (threshold =0.6)			Cosine coefficient (threshold =0.7)			TF*IDF weight(threshold=0.01)		
Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
1.33	50	2.59	13.34	58.82	21.74	38.67	36.25	37.42
75	1.17	2.3	67.85	1.06	2.08	60.71	6.67	12.01
38.165	25.585	2.445	63.33	7.2	11.91	49.69	21.46	24.71

Table 2: : result using 4-gram

c) N-gram (5 characters)

Dice's coefficient (threshold =0.6)			Cosine coefficient (threshold =0.7)			TF*IDF weight(threshold=0.01)		
Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
1.33	50	2.59	8	42.85	13.48	6.67	45.45	11.65
71.42	7.905	14.23	71.42	8.26	14.81	67.85	9.74	17.03
36.375	28.95	8.41	39.71	25.55	14.14	37.26	27.59	14.34

Table 3 : : result using 5-gram

Work to be done

- Evaluate techniques that use stemming with character n-grams
- Crowdsource an Arabic-English lexicon

References

- Al-Kharashi, Ibrahim A., and Martha W. Evens. "Comparing words, stems, and roots as index terms in an Arabic information retrieval system." *Journal of the American Society for Information Science* 45.8 (1994): 548-560.
- Chen, Aitao, and Fredric C. Gey. "Building an Arabic Stemmer for Information Retrieval." *TREC*. Vol. 2002. 2002.
- Gey, Fredric C., and Douglas W. Oard. "The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries." *TREC*. 2001.
- Larkey, Leah S., Lisa Ballesteros, and Margaret E. Connell. "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis." *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002.