Assessment Tools for Enhancing the Quality and Retrieval Efficiency of Arabic Web Content

by

Adnan Yahya

(joint work with Ali Salhi)

Birzeit University, Palestine

April 12, 2014
Outline

- Motivation and Introduction.
- Quality metrics.
- Similarity Measures
- Putting it together
- Next Steps.
Motivation and Introduction

Introduction

• Web content is increasing at a fast pace, more so for Arabic
• Content generated by humans, machines and jointly
• Still, Arabic is comparatively small relative to population size
• Large variance in quality: from Encyclopedia to Social Media
• A variety of language vehicles: form MSA to Dialects
• Lots of media: text, voice, pictures and video. We deal with text
• Subject to study by many, mainly in industrial nations (Googles, IBMs, BBNs and more)
• Much work for English but much less for Arabic (Why?)
Motivation and Introduction

• Given a text $T$, estimate its quality and make it known to the user

• Allow the user to access material in Arabic or other languages that may satisfy information need by returning material similar to need in multiple languages (without translation!)

• The user may opt to use the results to:
  – Improve article quality if current quality is below needed
  – Have access to good quality foreign material with a chance to translate
  – Detect duplicate material even Cross Languages (CL plagiarism detection)
  – Quality augmented/driven Information Retrieval (IR)
Motivation and Introduction

The story line

We have the following story line:

- For a search we need to return (high) **quality** content: we talk about how to measure text quality for Arabic content.
- The user may gain even if **only** low quality content is found: **less reliance on such content (a grain of salt!)**
- We may also need to return relevant info from other sources, even in other languages, need similarity checking: **how to measure the relatedness (semantic similarity) of two texts**
- So we may work with a single language or Cross Languages
Motivation and Introduction

The need for Automation

- Manual processing of content is out of the question due to SIZE
- So much can be gleaned from text, even when a human cannot see it! How does word usage change over time?
- Automation saves time and money, manual seed though!
- We need to quantify quality (have measures) and be able to detect similarity to ascertain that the found material is relevant
Motivation and Introduction

Some Relevant Properties of Arabic Writing

• Arabic is different in many ways: not all that is developed for other languages is applicable to Arabic as is.
• Consider: absence of capitalization, absence of diacritics, tolerance of spelling errors (say Hamza), coexistence with dialects; writing rules: one word sentences, lax punctuation, writing directionality, and more.
• However, it shares a medium size alphabet, better correspondence between the written and spoken, derivation rules, and more.
• So: Methods developed for other languages will need to be adapted to Arabic: a focus here!
Motivation and Introduction

The Wikipedia

Content/Article quality changes:
The Wikipedia (Arabic and other) used intensively. WHY?

– Well annotated: categorized, tagged, edited, with edit history and linked to similar material. **We use most of these features**
– Language is reasonable. Article quality is subject to discussion: so no uniform quality here **(feature, good, random)**
– Multiple authors, topics, editors: one can study this as well.
– Large and growing. Statistically sound: in Arabic 240K, in English 3500K and growing
– Good coverage also by topic
– Other resources can be used/added (WordNet, Dictionaries,...)
Quality Metrics
Quality in Wikipedia and General Texts

• What defines Quality:
  – **Language** parameters and style: simple/sophisticated, punctuation usage, sectioning, ...
  – **Contributor** Credibility: Author and Editor
  – **Supporting** materials: links (outbound and inbound), pictures, graphs,
  – **Currency**: updated when needed: though too many updates may mean “still developing” status
  – **Access** frequency and history

• A combination of all! But we don’t need to be that accurate!

• Recall: *Wikipedia* is highly annotated: including on quality: Feature(**gold**), Good (**silver**), Random (300,300, 240K)
Quality Metrics

Language

– General vs Specialized: can be determined by OOV words against a general (non-specialized) dictionary. Can use a general newswire corpus for the general dictionary.

– Some phrases/terms are pointers to good quality:
  • Despite, not withstanding, respectively, ....
  • بالرغم من ذلك، محض صدفة، قياسا على،
  • Stylistic issues like punctuation, sentence length, vocabulary count, ...

– The use of other languages (Monolinguality), including dialect

– Error Rate: ordinary and confusion letters (Hamza, Alef)

– Vocabulary: regular vs simple, regular vs children, ...

– Diacritics: total or partial: usually none
Quality Metrics

Links and Length

• Links:
  – Links are important in page ranking
  – Both inbound and outbound links are of value
  – Links to good pages: more weight than link to average pages

• Length:
  – Short articles are not as good as short
  – One may ignore pages of less than 40-50 words: can’t tell much in so many words
  – Different for other material (Multimedia)!
Quality Metrics
Contributors: Authors and Editors

• Edit History: Preserved Completely!
  – Temporal: how frequently changes occur, how much changes in each edit, what survives edits
  – What is the “Quality” of the edit author: good authors do good edits and produce good articles and good articles are produced by good authors/editors
  – Good authors/editors share networks: work on same articles. Working with a good author improves your reputation. Author credibility is affected by his/her network
  – A way to estimate quality is to credit each word by its author reputation, and to define author reputation by the quality of words he/she contributed: the process is iterative
  – Yes. It is a cycle. The process may be iterative!
Similarity Measures

Semantic vs Syntactic Similarity

• How semantically similar/related articles are (meaning!)
• Complicated by style, paraphrasing and synonyms
• Similar if they are telling the same story? Well almost: similar stories, related stories: a continuum from 0 to 1
• Categorization has an element of similarity
• But our concern: similarity between articles: single language or Cross Lingual (CL)
• Useful in plagiarism detection, IR: retrieve documents similar to the Information Need (Query)
• For us: find candidates for display, translation, relevant
Similarity Measures

Approaches to measuring Similarity

• **Bag of words**: distance tells how similar documents are. Problem: synsets, doesn’t work across languages; can’t detect similarity of summaries to original; or document to a query: length matters

• **Explicit Semantic Association (ESA)**:
  – Express texts in terms of *concepts*: a fixed number of concepts.
  – Each word is represented by a concept vector,
  – Each text is represented by the sum of its words concept vectors
  – Text chunks: similar if they have close enough concept vectors
  – Size irrelevant. problem: cross language difficulties.
  – Cross Language (CL) ESA: have common concepts (and vectors)

• **Wikipedia** can be the link!
Similiarity Measures

ESA:

• Each word is represented by a concept vector (of Wikipedia articles)
• Each text is represented by the sum of its word vectors
• Text size doesn’t matter: all texts map to a vector
• Similarity is judged by distance between the “text” vectors

CL-ESA:

• Consider only parallel articles in the two Wikipedia (e.g. Ar, En)
• Each word is represented by a concept vector: Wikipedia articles in OWN language: same dimensionality: comparable cross languages
• Again, each text is represented by the sum of its word vectors
• Similarity is judged by distance between the two vectors
• Need enough of credible parallel articles: (100,000?)
**Similarity Measures**

*Wikipedia* can be *the link*

- *Wikipedia* is the anchor link through its article words: generate an inverted table: for a word \( w \) associate \( n \)-dimensional vector \( V(w) \) with \( w \)-frequency in the \( n \) articles as elements. \( n \) is the Wikipedia Size!

  - In ESA Wikipedia *Articles* are the *concepts*
  - For CL ESA parallel articles alone are considered!

Vectors in *both* languages have *same* dimensionality

- The infrastructure exists: have enough *parallel* articles between Arabic and English (need not be limited to EN)
- We use categories/synsets: Wikipedia still the connection
- Measures of success: retrieving similar articles from the Wikipedia, or *close enough* ordering of similar articles
Similiarity Measures
ESA Example

The **man caught stealing** was sent to **jail** for **years**
The **thief** spent **long time in prison**

- **Thief** Vector= 9001007070100 Quite
- **Steal** Vector= 9000107081100 Similar
- **Prison** Vector= 7000004080100 Quite
- **Jail** Vector= 7001105070100 Similar
- **Time** Vector= 1001807161200 Quite
- **Years** Vector= 0000806081100 Similar
- **Word frequencies count**
- Imagine summing for both sentences: the sums (averages) should be close. The numbers represent the Concepts (articles, categories)
- Imagine the sentences in different languages: matters little (just limit vectors to parallel articles)
Putting it Together

• The goal is to improve the quality of Arabic Web Content
• We evaluate current content and tag it and offer people the chance to improve
• When we have a better quality foreign article we offer it as a possible source and a translation candidate
• Text size independence allows the process to start from the specification of user information need (query)
• We can even offer possible terms/words for inclusion in a new/improved Arabic article
• Results apply to other language pairs with infrastructure
• One potential applications: Plagiarism Detection
Next Steps

• Done some testing but much more needs to be done
• So far, more results on Wikipedia Article quality and less on similarity measures: that’s the focus
• The integration of the components is as important
• Extension to other types of texts including short posts or user need specifications: we want to be able to move from a query (or a query stream) to the suggestion of translation (Foreign) / improvement (Arabic) articles
• The tools don’t require deep understanding, though understanding helps developing heuristics and fine-tuning
• The good part is: mostly automated
Thanks