



Upshot from Dayton: Found in Translation

Ranking Arabic Documents with
Machine Learning

Conops

- Understand the value of the document
- Limited Interpreters - upshoot documents that have urgent relevance
- Mobile app translation is nice to have



[14] أسرى آل الفرج والذات لعدم اجتماعهم فعلموا بأمر حله وسامحهم به بأي كانت الأولى في هذه الأوقات أن يرسل من يجه.

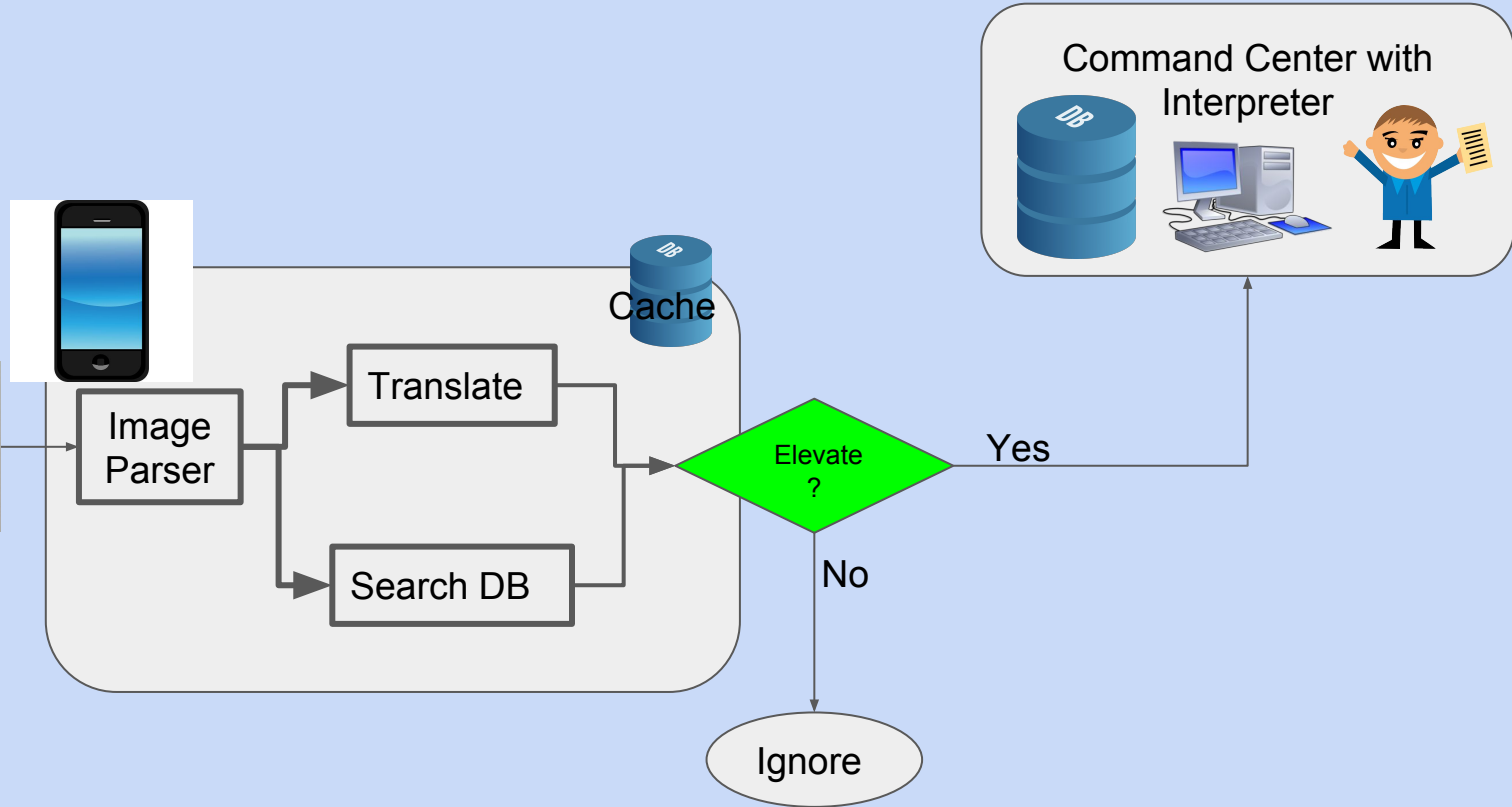
[15] أسير بعض آل العنصرين بسبب اجتماعهم مع أبيهم هو بنو عليم وقيل عوفون وقيل عسلة

صالح 383.

[16] بال صلوة وأحواله بعد عزمه على التبرؤ اليه بعد أن تم اكتشافه وقد عفا عنه معلوماً لغرضي والذاتي.

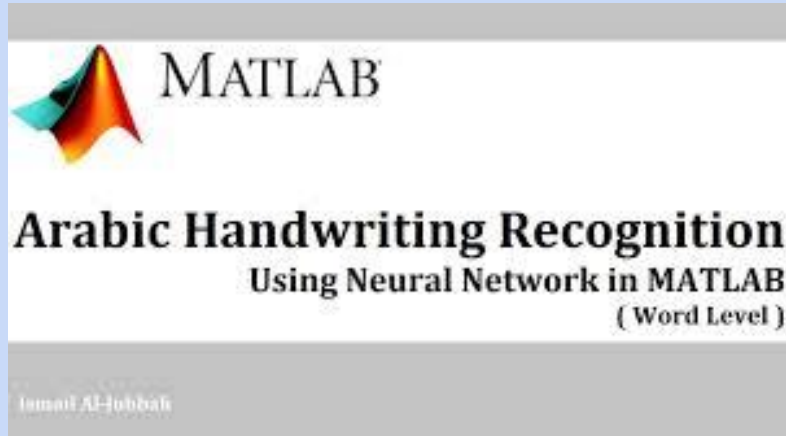
[17] في ذلك الموضع المذكور وكذلك كان لعده أجداد عديدة ما عداه من ذيا الباكستان خصوصاً وفضل حتى جاء المصطفى

[18] أسير وعظيم المخلص من الأجداد في البكوات والذات السامع للأخوة وبعضهم لغرض الردة عن حركتهم ومن كل مرة تأسلم.

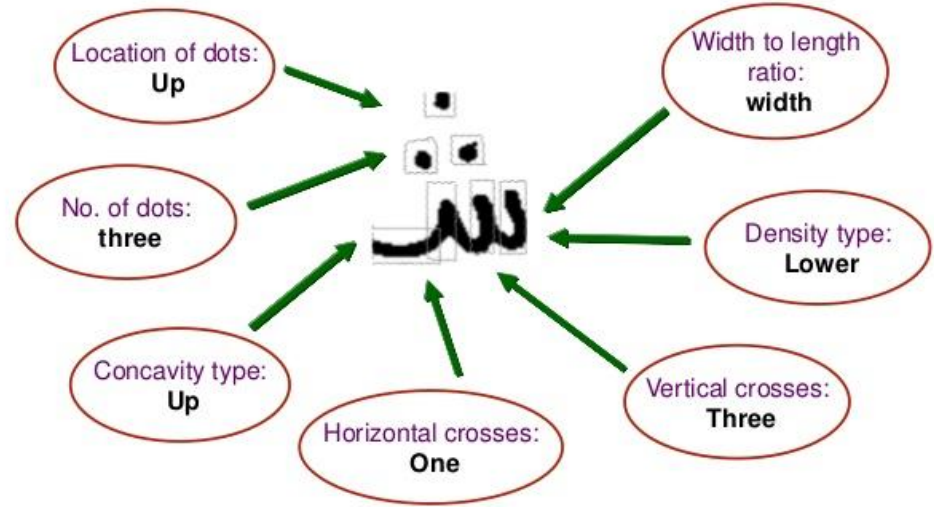


User Interface

Demonstrate Image Parser



Optical features of Arabic characters



Problem Statement

- Given a document photo, what other documents is it like?
- Should someone translate it?

صدى 10/13 pt

من عداد الصنائع الإنسانية وهو رسوم وأشكال حرفية تدل على الكلمات المسموعة الدالة على ما في النفس. فهو ثاني رتبة عن الدلالة اللغوية وهو صناعة شريفة إذ الكتابة من خواص الإنسان التي يميز بها عن الحيوان. وأيضاً فهي تطلع على ما في الضمائر وتتأذى بها الأغراض إلى البلد البعيد فتقضى الحاجات وقد دفعت مؤونة المباشرة لها ويطلع بها على العلوم والمعارف وصحف الأولين وما كتبوه في علومهم وأخبارهم فهي شريفة بجميع هذه الوجوه والمنافع.... فالخط المجرد كماله أن تكون دلالته واضحة بإبانة حروفه المتواضعة وإجادة وضعها ورسمها كل واحد على حدة متميز عن الآخر إلّا ما اصطلاح عليه الكتاب من إيصال حرف

10/13 pt Seria Italic

Writing is the outlining and shaping of letters to indicate audible words which, in turn, indicate what is in the soul. It comes second after oral expression. It is a noble craft, since it is one of the special qualities of man by which he distinguishes himself from the animals. Furthermore, it reveals what is in (people's) minds. It enables the intention (of a person) to be carried to distant places.... It enables (people) to become acquainted with science, learning, with the books of the ancients, and with the sciences and information written down by them. Because of all these useful aspects, writing is a

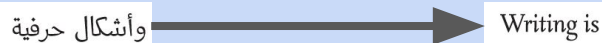
Limitations and Assumption

Limitations:

- Small time and memory constraints (mobile computer)
- Limited connectivity to central database
- Limited central intelligence about documents in other language

Assume:

- Documents can be parsed into words
- Words can be mapped to their English word/words (Dictionary lookup)
- English words can be counted to create a histogram of words (BOW)



Demonstrate Classifier Solution

Key Idea:

- Convert document to a simple representation in native language
 - Represent documents as bag of words (BOW) histograms
- This allows train a classifier(s) using corpus of labeled documents in native language **(HUGE CORPUS FROM NATIVE INTELLIGENCE)**
- Classifiers quickly tell what kind of document it is like.
 - Optionally, also classify as “Important for further detailed translation”

Benefits:

- Simple classifiers on BOW histograms are just linear feature vectors
 - Low memory, ultra fast decisions
- As more intel is gained, the decision rule(s) can be updated

What Classifier?

- Classifier 1: Is it important for detailed translation by human?
- Classifier 2: What type of document is it like?/Which documents that I know are like it?
- You want:
 - Decision confidence
 - Fast lightweight decision
 - Small data decision update
 - **Nearest neighbor:** Distance implicitly tells you confidence about what type of document it is
 - **Relevance vector machine (RVM):** Gives you the category as well as explicit confidence score $[0,1]$
Simple weight vector **(super small memory)**

Limitations

Hidden assumption:

- BOW for word -> word translation will give proper representation for classifying using our English corpus.

If not:

- Learn an intermediate mapping that re-weights other language BOW to native language BOW.
 - Simple least squares reweighting of histogram
 - Cross correlational analysis

Summary

List of Resources

- <http://arxiv.org/pdf/1206.1518.pdf> - research paper for recognizing Arabic characters
- <http://www.amara.org/en/videos/cl4VRKdprGv/info/arabic-handwriting-recognition-using-neural-network-in-matlab-word-level-user-manual/> - video tutorial to use Matlab's Arabic Handwriting Recognition feature
- <https://console.ng.bluemix.net/catalog/services/visual-recognition/> - specs for Watson's Visual Recognition service
-

Backup Slides

ضاع في الترجمة

May 22, 2016

Safe Sign Group

Language Recognition

- There are thousands of languages spoken on Earth, but the top ten languages cover most speech.
- In field operations, English would be expected to be the default language in only a small percentage of situations.
- Arabic is spoken by

Modern Standard Arabic

- Challenges of translating content, and also obtaining the right context.

Simple Yet Complicated

