CREATING RESOURCES FOR ARABIC SUMMARISATION

Dr Mahmoud El-Haj

School of Computing and Communications

PURPOSE

The purpose of this exercise was two-fold:

- First it addresses a shortage of relevant data for Arabic natural language processing.
- Second it demonstrates the application of crowd-sourcing, Machine Translation and Human Experts to the problem of creating natural language resources.

Creating Resources for Single Document Summarisation

- Contribution
  - Provide a human single-document summaries corpus.
- Method
  - We used online workforce service (Mechanical Turk).
- Source
  - Arabic Wikipedia and two Arabic newspapers (Alrai, Alwatan).

The Document Collection

The sources were chosen for the following reasons.

- Contain real text written and used by native Arabic speakers.
- Written by many authors from different backgrounds.
- Cover a range of topics from different subject areas.

HUMAN-GENERATED SUMMARIES

- The corpus was generated using Mechanical Turk
  - Human Intelligence Tasks (HITS).
  - The assessors (workers) summarise one article per task.
  - No more than half of the sentences in the article.
  - Five summaries were created for each article.
  - Summaries for a given article were generated by different workers.

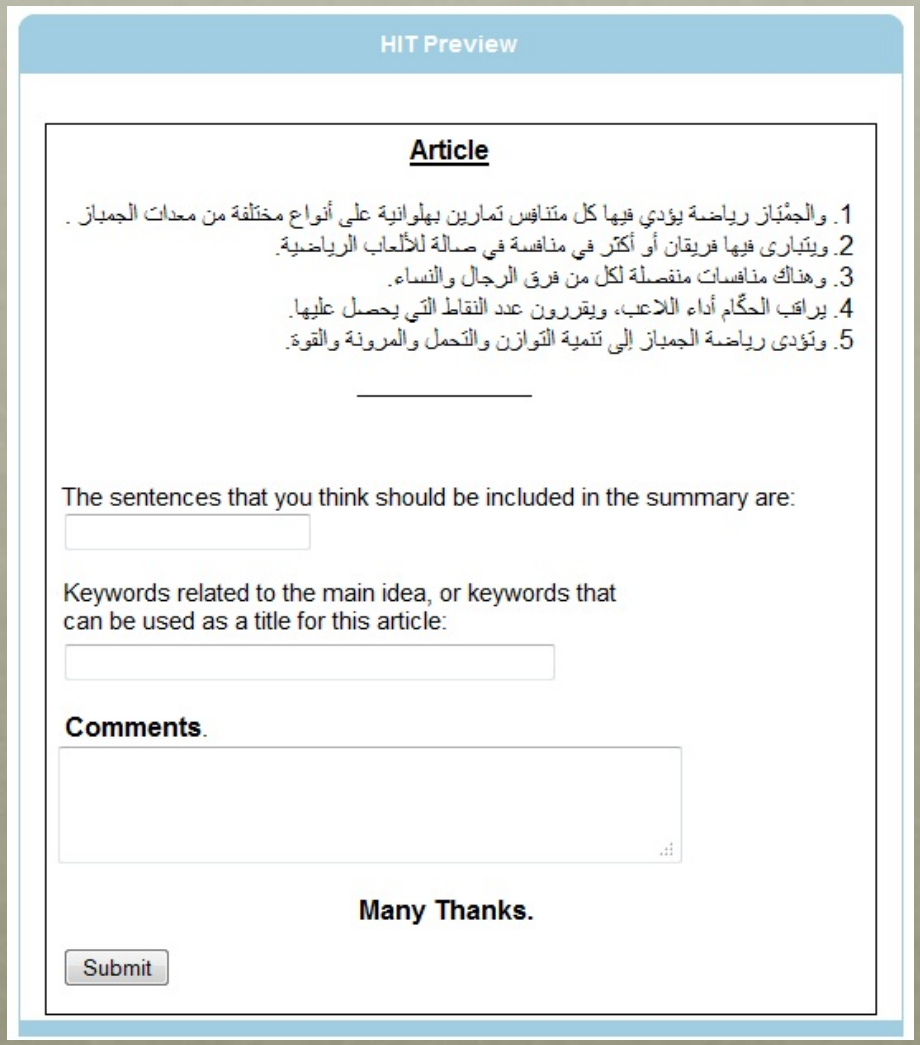
Quality of the Summaries

- Each worker was asked to provide up to three keywords.
- In the case of selecting random sentences, the summary is still considered.

Creating Gold-Standards

- First: We selected all sentences identified by at least three of the five annotators (Level 3).
- Second: We selected all sentences identified by at least two annotators (Level 2).
- Finally, all sentences identified by any of the annotators (All).

HIT SAMPLE


HIT Preview
Article
1. والجمْبَاز رياضة يؤدي فيها كل متنافس تمارين بهلوانية على أنواع مختلفة من معدات الجمباز .
2. ويتبارى فيها فريقان أو أكثر في منافسة في صالة للألعاب الرياضية.
3. وهناك منافسات منفصلة لكل من فرق الرجال والنساء.
4. يراقب الحكّام أداء اللاعب، ويقررون عدد النقاط التي يحصل عليها.
5. وتؤدى رياضة الجمباز إلى تنمية التوازن والتحمل والمرونة والقوة.
The sentences that you think should be included in the summary are:
Keywords related to the main idea, or keywords that can be used as a title for this article:
Comments.
Many Thanks.
Submit

EASC CORPUS STATISTICS

Corpus Name	Essex Arabic Summaries Corpus (EASC)
Number of Documents	153
Number of Sentences	2,360
Number of Words	41,493
Number of Distinct Words	18,264
Number of Human Summaries	765 (five for each document).

Creating Resources for Multi-Document Summarisation

- Contribution
  - Provide a human and system multi-document summaries corpus.
- Method
  - We used Machine Translation and Human Expert (Arabic Native Speakers).

USING MACHINE TRANSLATION

- MT was used to overcome the lack of Arabic multi-document resources (gold-standard summaries).
- The DUC–2002 English dataset provides articles and multi-document gold standards for extractive summaries.
- We performed sentence-by-sentence translation of this dataset into Arabic.
- The translation is to allow us to evaluate Arabic summaries against English gold standards.
- The translating into Arabic was done using Google Translate.

USING MACHINE TRANSLATION

- Main objective is to compare our Arabic multi-document summariser with the currently available English ones.
- To do this we translated the sentences in each article into Arabic, using the Java version of the Google Translate API.
- A total of 17,340 sentences were translated.
- The process resulted in a parallel sentence-by-sentence Arabic/English version of the DUC-2002 dataset.
- The summaries were created using the applicable English or Arabic version of our multi-document summariser system.
- We did not translate the gold-standard summaries themselves into Arabic.

DUC-2002 ARABIC CORPUS STATISTICS

Corpus Name	Duc-2002 (Arabic Translation)
Number of Documents	567
Number of Sentences	17,340
Number of Words	199,423
Number of Distinct Words	19,307
Number of Reference Sets	59
Documents per Reference Set	10 on average
Number of Gold-standard summaries	118 (two for each reference set)

USING HUMAN EXPERTS

- Used human participants to manually create the resources.
- The manual process included translating, validating and summarising documents written in English.
- The process of manually creating an Arabic multi-document corpus was part of our organisation for the TAC–2011 MultiLing Summarisation Pilot.

USING HUMAN EXPERTS

- The participants are responsible for translating and summarising an English test collection into six languages including: Arabic, Hindi, French, Czech, Greek and Hebrew.
- The test collection was based on WikiNews texts.
- The source documents contained no meta-data or tags and were represented as UTF–8 plain text files.
- The test collection contained 100 articles divided into ten reference sets, each contained ten related articles discussing the same topic.
- The original language of the dataset was English.

THE PARTICIPANTS

- A total of twelve people participated in:
  - translating the English corpus into Arabic.
  - summarising the set of related Arabic articles.
  - validating the translation.
  - evaluating the summarisation quality.
- The participants were paid using Amazon vouchers.
- The amount of the vouchers varied depending on the task performed.
- The total amount of Amazon vouchers paid to the participants was £250 where three of the participants volunteered to do the tasks.

THE PARTICIPANTS

- The participants have been selected based on their proficiency in Arabic and each one of them must be an Arabic native speaker.
- The participants are studying, or have finished a university degree in an Arabic speaking country.
- The participants age range between 22 and 64 years old.

CREATING THE CORPUS

- The participants translated the English dataset into Arabic.
- For each translated article another translator should validate the translation and fix any errors.
- For each of the translated articles, a number of three manual summaries were created by three different participants (human peers).
- Amid the summarisation process the participants should evaluate the quality of the generated summary by assigning a score between one (unreadable summary) and five (fluent and readable summary).
- No self evaluation was allowed.

TAC–2011 ARABIC CORPUS STATISTICS

Corpus Name	Duc-2002 (Arabic Translation)
Number of Documents	100
Number of Sentences	1,573
Number of Words	30,908
Number of Distinct Words	9,632
Number of Reference Sets	10
Documents per Reference Set	10
Number of Gold-standard summaries	30 (three for each reference set)

SUMMARY

- Resource creation plays an important role in the advance of Arabic single and multidocument summarisation.
- Three summaries corpora have been created.

1. Essex Arabic Summaries Corpus (EASC), a single-document Arabic extractive summaries.
2. The second corpus was a multidocument summaries created using a machine translation tool. The approach used introduced a cost-effective solution to the problem of limited Arabic resources.
3. The third corpus, MultiLing dataset, is a multidocument summaries corpus. The corpus was created by native Arabic speakers.

http://www.lancs.ac.uk/staff/elhaj/