
# Towards Building a Standard Dataset for Arabic Keyphrase Extraction Evaluation

Muhammad Helmy, Marco Basaldella, Eddy Maddalena, Stefano Mizzaro and Gianluca Demartini

Artificial Intelligence Laboratory @ University of Udine - http://ailab.uniud.it

# Keyphrase Extraction (KPE)

- Keyphrases (KPs) are short phrases that best represent a document content
- They can be useful in a variety of applications, including document summarization, classification, indexing and retrieval models
- Automatic KPE has two approaches to identify KPs of a document: supervised and unsupervised
- supervised approach requires a training dataset for its machine learning algorithm and both of them require a gold standard dataset to evaluate the extracted KPs

# Dataset for KPE

- Many datasets have been proposed in the past years, prominently for the English language
- All the well-formed datasets available cover only English. This fact obviously hinders the development of a multi-lingual KPE community
- There is a growing interest around the problem of KPE in the Arabic language.
- Arabic is, in fact, the fifth most spoken language in the world, with more than 330 million native speakers and growing digital content
- There is no shared, standard dataset that researchers can use to assess the performance of their algorithms

# Current state of Arabic Datasets for KPE

- The researchers are using custom-made corpora in their work, remarking the absence of a standard dataset
- Training and evaluation are performed over manually annotated documents including news and Wikipedia articles and their meta-tags
- Most of the documents are annotated by the researchers themselves, which creates some bias
- The aim of this work is then to provide a dataset for the growing community of Arabic KPE by mean of Crowdsourcing to train and evaluate their KPE algorithms

# Dataset Development Roadmap

- The Crowdsourcing Module
  - *Document Collection*
  - *Task Design*
  - *Keyphrase Collection*
  - *Descriptive Statistics*
- The Keyphrase Collection Module
  - *Selecting High Quality KPs*
  - *Data Validation*
  - *Applying a Baseline KPE System on the Dataset*

# The Crowdsourcing Module

## *Document Collection*

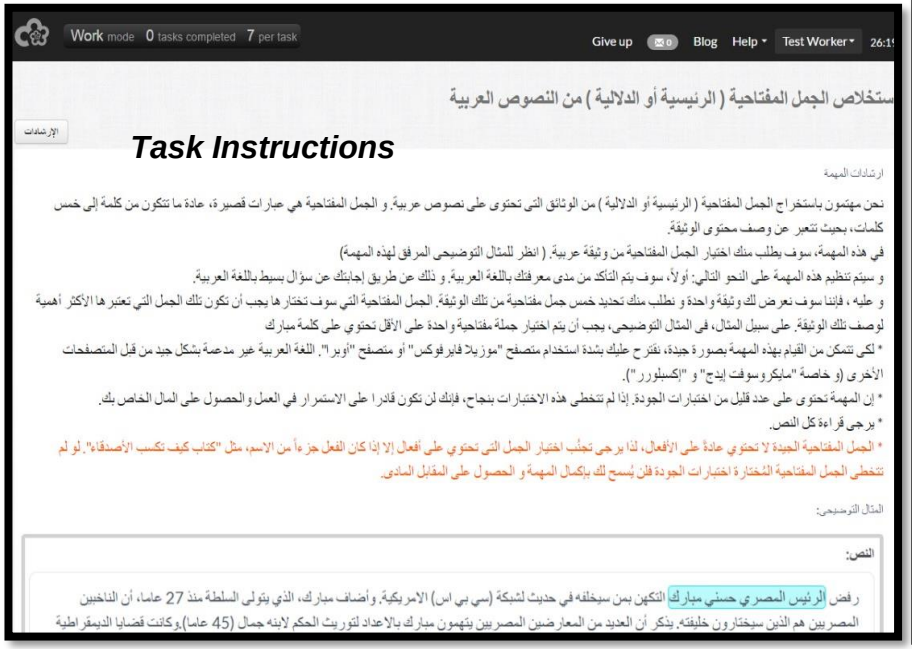
- The collection contains 160 documents selected from four general purpose, freely available corpora and categorized into nine topics
- the documents have been preprocessed, by eliminating unrelated text like headers, image captions, and corpus metadata
- Document length vary between 500 and1000 words, with a median of 735.5

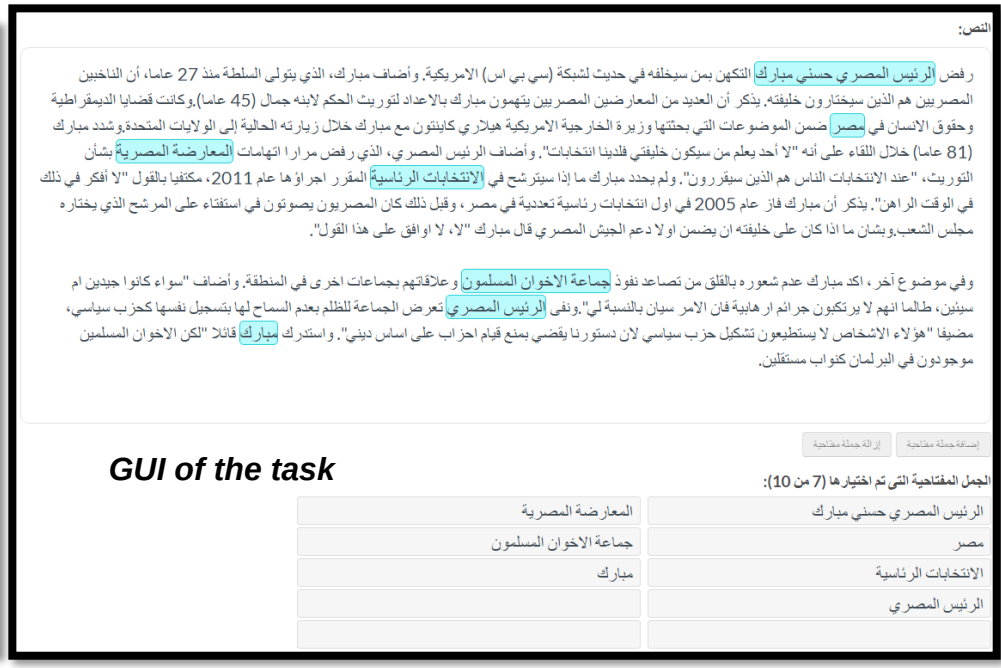
| Category | Number of documents |
|---|---|
| ArtAndMusic | 18 |
| Environment | 18 |
| Finance | 18 |
| HealthAndMedicine | 19 |
| Politics | 18 |
| Religion | 17 |
| ScienceAndTechnology | 18 |
| Sport | 17 |
| Tourism | 17 |
| **Total:** | **160** |

| Corpus | Number of documents |
|---|---|
| Essex Arabic Summaries Corpus (EASC) | 31 |
| Arabic Newspapers Corpus (ANC) | 46 |
| Corpus of Contemporary Arabic (CCA) | 53 |
| Open Source Arabic Corpora (OSAC) | 30 |
| **Total:** | **160** |

# The Crowdsourcing Module

## *Task Design*




# The Crowdsourcing Module

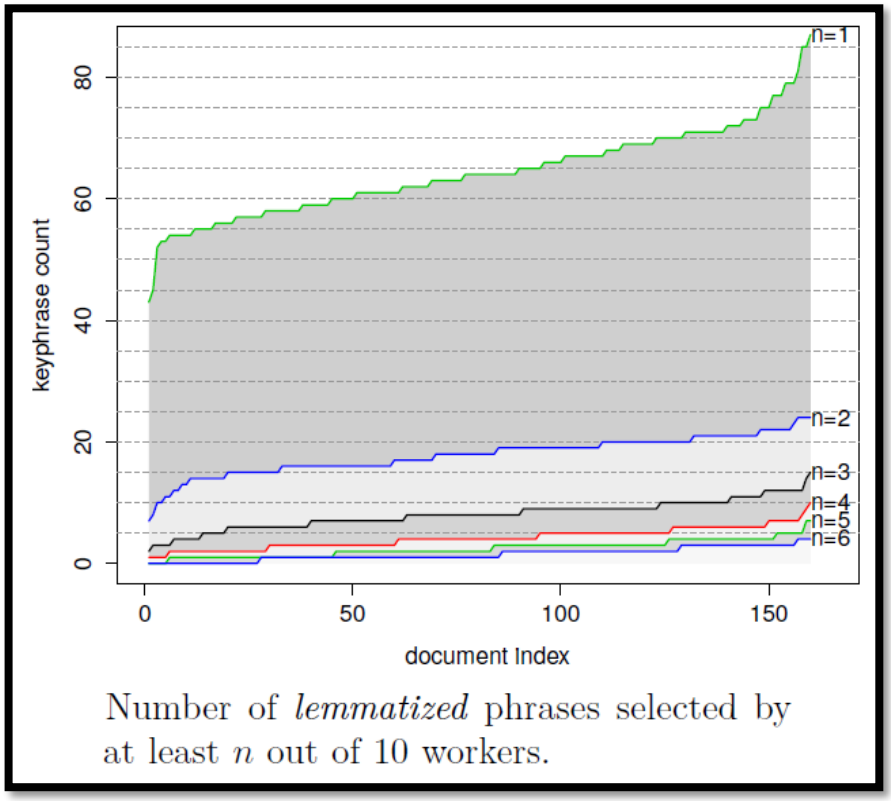
## *Keyphrase Collection*

- A pilot experiment was launched on Crowdflower platform with 10 documents, to tune the task for the whole corpus
- We decided to use 10 workers per document, and to ask each worker to select 10 KPs, while in the pilot experiment we required just 5 KPs by 5 workers
- Moreover, we adjusted the task instructions to guide workers to not select phrases beginning with stopwords, verbs, or adjectives.
- Any unit was discarded if the worker did not spent at least 120 seconds on the document
- Each worker could read and annotate up to ten documents
- Finally, we required Crowdflower to select only medium and highest quality workers

# The Crowdsourcing Module

## *Descriptive Statistics*

- The experiment was completed by a total of 226 workers, for a mean of 7.07 documents per worker
- More than 75% of the workers were based in one of four countries, namely Egypt, Algeria, Saudi Arabia and Tunisia
- Only 2.2% of the workers came from countries where Arabic is not an official language, i.e., Germany, Indonesia, Netherlands, France, and Turkey



Number of *lemmatized* phrases selected by at least $n$ out of 10 workers.

- The time spent reading a document had an average of 302 seconds (5 minutes) and a median of 222 seconds (less than 4 minutes)
- We collected a total of 10’646 distinct KPs

# The Keyphrase Collection Module

## *Selecting High Quality Keyphrases*

- To rank the extracted KP, we adopted two selection approaches
- *Frequentist*: we order KPs by the number of times that they have been selected by workers, then we discard all the KPs that have not been selected at least twice
- *Linguistic*: we build a language model and sort the KPs using that model; then, we keep the best 15 ranked phrases per document and discard the others

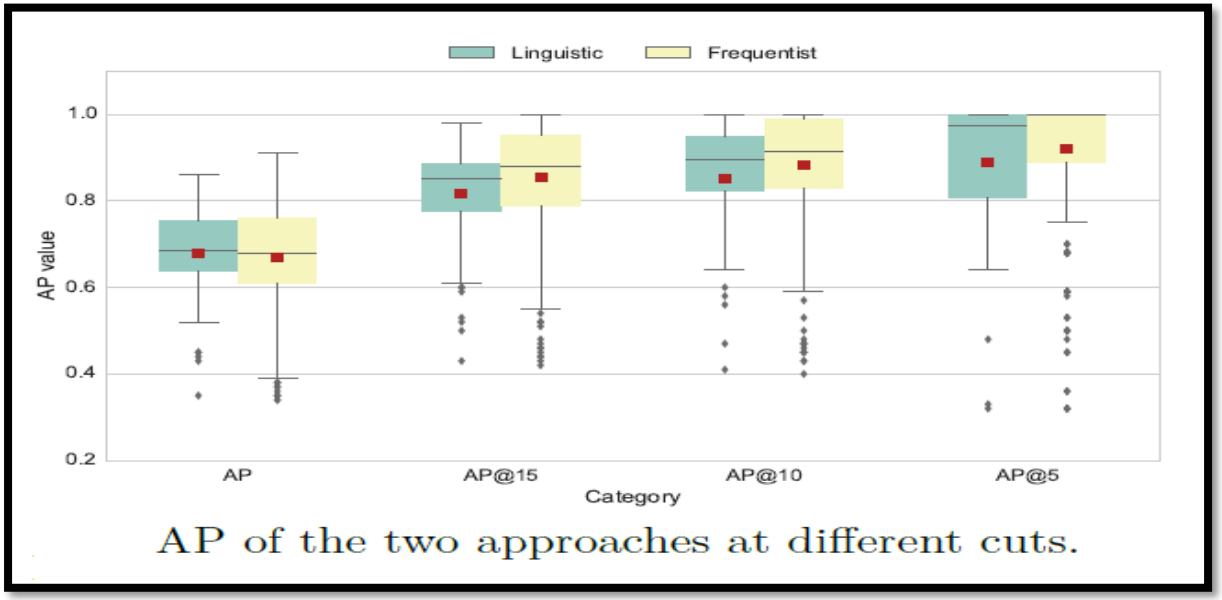
Number of KPs which have been selected by at least $n$ crowd workers for each document.

| $n$ Worker | Median | Mean | Min | Max |
|---|---|---|---|---|
| 2: | 18 | 17.8 | 10 | 24 |
| 4: | 4 | 4.1 | 1 | 10 |
| 6: | 1 | 1.5 | 0 | 4 |
| 8: | 0 | 0.5 | 0 | 2 |
| 10: | 0 | 0.1 | 0 | 1 |
| Linguistic cut: | 15 | 15.5 | 15 | 19 |
| SEMEVAL: | 14 | 15.1 | 8 | 37 |

# The Keyphrase Collection Module

## *Data Validation*

- To validate our approaches, we selected a subset of 56 documents from the corpus and had an expert (an Arabic native speaker doing a PhD on KPE) manually assess the quality of the KPs that the crowd selected
- The expert was shown the KPs in random order to avoid any bias
- We use the classical ones: Average Precision (AP) and Mean AP (MAP), as well as MAP@5, MAP@10, and MAP@15 to show the quality of the first ranked KPs




AP of the two approaches at different cuts.

# The Keyphrase Collection Module

***Applying a Baseline KPE System on the Dataset***

- Various KPE systems employ TF-IDF as a numerical and statistical method to extract KPs
- An Arabic TF-IDF based testbed system was implemented as a baseline KPE to evaluate the quality of the dataset KPs and assess workers performance
- For each document, two lists of words have been generated. The first list contains words of all KPs extracted by the workers excluding stopwords while the second one is a sorted list of the important words generated and ranked by the testbed system
- After that, the two lists were compared and the precision of the dataset was calculated
- The precision was about 0.6 which means that 60% of workers KPs words are recognized by the system;

# Future Work

- Being our first effort in building such a corpus, there is plenty of directions to explore in the future
- It is possible that we will enlarge the corpus by including more documents
- Try different approaches and variants to filter the high quality KPs
- It will also be important to understand which is the ideal number of workers per document
- We also plan to try different experimental designs. For instance it would be interesting to try an approach similar to the well known ESP game, including the mechanism of taboo words to avoid the crowd to repeatedly select already known KPs
- The dataset is available at https://github.com/ailab-uniud/akec

شكرا
Thank you