“KALIMAT” a Multipurpose Arabic Corpus

Mahmoud El-Haj
Lancaster University
m.el-haj@lancaster.ac.uk

Rim Koulali
Mohammed 1 University
rim.koulali@gmail.com

KALIMAT

LANCASTER UNIVERSITY

- KALIMAT (Arabic transliteration of “WORDS”).
- Resources, such as corpora, are important for researchers working on Arabic Natural Language Processing (NLP)

Motivation

LANCASTER UNIVERSITY

- Shortage of Arabic resources.
- Lack of Arab participants to create such resources.
- Cost (time) of manual corpus creation.
- Lack of standardisation.
- Benefit from current Arabic NLP tools.
- Use the annotated and summarised corpus as baseline.

KALIMAT Statistics

LANCASTER UNIVERSITY

# ARTICLES	KALIMAT
20,291	Arabic articles (Abbas et al. 2011)
20,291	Extractive single-document system summaries
2,057	Multi-document system summaries
20,291	Named Entity Recognised articles
20,291	Part of speech tagged articles
20,291	Morphologically analysed articles

Collection Categories

LANCASTER UNIVERSITY

Topic	Number-of-Articles	Number-of-Words
Culture	2,782	1,359,210
Economy	3,468	3,122,565
International News	2,035	855,945
Local news	3,596	1,460,462
Religion	3,860	1,555,635
Sports	4,550	9,813,366

KALIMAT Creation Process

LANCASTER UNIVERSITY

- The process of creating KALIMAT was applied to the entire data collection (20,291 articles).
- We used Arabic NLP tools from the literature.
- The reason behind selecting these tools:
  - trained and tested on actual Arabic datasets.
  - tuned to provide high quality results.

LANCASTER UNIVERSITY

1- Arabic Automatic Summarisation

Single and Multi

a- Single-document Summarisation

LANCASTER UNIVERSITY

- Gen-Summ (El-Haj et al. 2010) is a single document summariser based on VSM model (Salton et al. 1975)
- Takes an Arabic document and its first sentence and returns an extractive summary.
- A number of 20,291 system summaries have been generated.

b- Multi-document Summrisation

LANCASTER UNIVERSITY

- Cluster-based (El-Haj et al. 2011) is a multi-document summariser that treats all documents to be summarised as a single bag of sentences.
- The sentences of all the documents are clustered using different number of clusters.
- A summary is created by selecting sentences from the biggest cluster only (if there are two we select the first biggest cluster).
- We generated 2,057 multi-document summaries.
- With a summary for each 10, 100 and 500 articles in each category, in addition to a summary for all the articles in each category.

Multi-document Summaries

LANCASTER UNIVERSITY

Topic	10	100	500	all	Total
Culture	250	25	5	1	281
Economy	327	33	7	1	368
International News	169	17	4	1	191
Local news	324	33	7	1	365
Religion	348	35	7	1	391
Sports	410	41	9	1	461
					2,057

LANCASTER UNIVERSITY

2- Named Entity Recognition (NER)

Arabic Named Entity Recognition

LANCASTER UNIVERSITY

- We used an Arabic Named Entity Recognition system (ANER) (Koulali and Meziane 2012) to annotate the data collection.
- ANER was developed using dependent and independent binary features and SVM implementation for sequence tagging based on HMM.

NER Annotation Process

LANCASTER UNIVERSITY

- To annotate the data collection we followed the Computational Natural Language Learning (CoNLL) 2002 and 2003 shared tasks
- formed by tags falling into any of the following four categories:
  - Person Names: محمود درويش (Mahmoud Darwish).
  - Location Names: المغرب (Morocco).
  - Organisation Names: الأمم المتحدة (United Nations).
  - Miscellaneous Names: NEs not belonging to any of the previous classes(date, time, number, measurement, percentages).

Training ANER System

LANCASTER UNIVERSITY

- ANER was trained using ANERCorpus (Benajiba et al. 2007), a manually annotated corpus following the CoNLL shared task.
- The reason behind choosing ANERCorpus to train the system:
  - corpus articles were chosen from Arabic newswires and Wikipedia Arabic, which is quite close to Alwatan's data collection

IBO2 Annotation

LANCASTER UNIVERSITY

- ANERCorpus contains more than 150,000 tokens tagged according to the IBO2 annotation:
- B-PERS: the beginning of a person name.
- I-PERS: the continuation (inside) of a person name.
- B-LOC: the beginning of a location name.
- I-LOC: the inside of a location name.
- B-ORG: the beginning of an organisation name.
- I-ORG: the inside of an organisation name.
- B-MISC: the beginning of the name of an entity which does not belong to any of the previous classes (Miscellaneous).
- I-MISC: the inside of the name of an entity which does not belong to any of the previous classes.
- O: The word is not a named entity (Other).
- A percentage of 90% of the ANERCorpus was used for training and the remaining 10% was used for testing.

ANER Output

LANCASTER UNIVERSITY

- We used the ANER system to generate 20,291 NER annotated documents following IBO2 annotation.
- The annotated data collection could benefit researchers working on the Information Extraction, Question Answering and Machine Translation.

LANCASTER UNIVERSITY

3- Part of Speech Tagging

Stanford POSTagger

LANCASTER UNIVERSITY

- We used Stanford POSTagger (Toutanova et al. 2003) to annotate the 20,291 document collection.
- The strength of the Stanford POSTagger relies on the following points:
  - Explicit use of both preceding and following tag contexts via a dependency network representation.
  - Broad use of lexical features, including jointly conditioning on multiple consecutive words.
  - Effective use of priors in conditional log-linear models.
  - Fine-grained modelling of unknown word features.

Stanford POSTagger

LANCASTER UNIVERSITY

- A supervised system depending on different trained models.
- Arabic model was trained using the Arabic Tree-bank p1-3 corpus.
- The POStagger identifies 33 part of speeches, using the Penn Treebank project codification such as: Noun (NN), Plural Noun (NNS), Proper Noun (NNP), Verb (VB), Adjective (JJ).
- The tagger reached an accuracy of 96.50%.
- The POST annotated 20,291 documents could help researchers working on Arabic IR, Word Sense Disambiguation and supervised learning systems.

LANCASTER UNIVERSITY

4- Morphological Analysis

Alkhalil Morphological Analyser

LANCASTER UNIVERSITY

- Used Alkhalil morphological analyser (Mazroui et al. 2011).
- Alkhalil was written in Java, the lexical resources consist of several classes, each representing a type of the same nature and morphological features.
- The Analysis was carried out in the following steps:
  - pre-processing (removal of diacritics)
  - segmentation (each word is considered as [proclitic + stem + enclitic]).

Alkhalil Morphological Analyser

LANCASTER UNIVERSITY

- Alkhalil identifies possible solutions of the segmented words using their morphosyntactic features (i.e. vowelisation, nature of the-word, vowelled patterns, stems, roots, suffixes, prefixes and syntactic-forms).
- Applying Alkhalil analyser on the data collection we reached an accuracy of 96%.
- The morphological analysis of 20,291 documents could help in improving the performance of many tools such as: automatic vocalization, spell checking, automatic summarization.

LANCASTER UNIVERSITY

KALIMAT Output Samples

Text Sample

سالم الرحبي : تنطلق اليوم الدورة البرامجية الجديدة للتليفزيون والاذاعة وبرنامج الشباب والتي تستمر طوال اشهر ابريل ومايو ويونيو وتحمل في طياتها العديد من البرامج الجديدة والفقرات الشيقة التي تتناسب مع اذواق جميع المشاهدين والمستمعين على حد سواء . دورة البرامج الحالية راعى فيها المسؤولون في وزارة الاعلام التنوع والتجديد في البرامج اضافة الى مراعاة اوقات المشاهدين والمستمعين بجميع فئاتهم حيث تم الاعداد المسبق لخارطة التليفزيون بشكل منهجي من خلال نوعيات منتقاة من البرامج كما تم تعديل تشكيلة السهرات الاسبوعية وتغيير جدول البرامج الوثائقية بحيث تشمل التنوع الثقافي مع التركيز على طرح البرامج المحلية التجديد فيها . وقد اكدت ادارة التليفزيون في اللقاء الصحفي الذي عقدته ظهر امس بمكتب مدير عام التليفزيون المهندس عبدالله العبري وبوجود صالح بن محفوظ القاسمي مدير البرامج العامة بالتليفزيون وزوينة الراشدي منسقة مكتبة التليفزيون ان الادارة سعت جاهدة اجل الخروج بدورة متميزة تتماشى مع رغبات المشاهد العماني بالدرجة الاولى مع التركيز ايضا على المنافسة الصحية بين باقي القنوات الفضائية مشيرين الى ان ادارة التليفزيون اصبحت تختار البرامج التي تريد ان تطرحها في الدورة البرامجية بعد ان كانت تفرض بعض البرامج وجودها وذلك من اجل ارضاء المشاهد والخروج بالصورة اللائقة أمامه.

Figure 1: Data Collection Text Sample

Single-document Summary

دورة البرامج الحالية راعى فيها المسؤولون في وزارة الاعلام التنوع والتجديد في البرامج اضافة الى مراعاة اوقات المشاهدين والمستمعين بجميع فئاتهم حيث تم الاعداد المسبق لخارطة التليفزيون بشكل منهجي من خلال نوعيات منتقاة من البرامج كما تم تعديل تشكيلة السهرات الاسبوعية وتغيير جدول البرامج الوثائقية بحيث تشمل التنوع الثقافي مع التركيز على طرح البرامج المحلية التجديد فيها .

Figure 2: Single-document Summary Text Sample

Multi-document Summary

القاهرة (الوطن) : الحوار مع ممدوح عدوان ليس بحاجة لأي مقدمة فهذا الشاعر والمسرحي والروائي والسيناريست والمترجم السوري حالة لا تقبل الارتهان لأي سلطة أو قاعدة سوى ما يقوله في هذا الحوار عن الخضوع للإنسانية والإبداع كحالتين مطلقتين تحكمان حياته ! ! ممدوح عدوان في هذا الحوار الذي يجيء عفويا وصادقا يناكف الحقائق الراسخة في قعر الوعي ويمضي في التاريخ الشخصي بعيدا وفي زوايا غير معروفة ومن وضعه الصحي إلى العديد من التفاصيل في المشهد الثقافي إلى متاهة المبدع بين الأشكال الفنية المتعددة في البداية أود أن أسألك عن وضعك الصحي وهو ما يشغل الكثيرين من قرائك في البداية سأوجز عن حكاية المرض : ففي بداية عام 2003 بدأت أحس بتغيرات غير صحية أو تغيرات مزاجية في طبعي فمثلا أنا في العادة أحكي كثيرا وأضحك كثيرا ولم أعد أضحك أو أحكي أو بالأحرى فقدت شيئا من حيويتي وبعدها سافرت إلى القاهرة وعدت وكانت هناك ملاحظات من الأصدقاء علي يقولون فيها : ما به ممدوح وأنا لم أحس بشيء غير طبيعي في وحينما عدت للعمل لاحظت أنني بدأت أنسي بشكل غير طبيعي فعندما كنت أكتب حوارية ما بين شخصين كنت أنسى أحدهما ! أو حين كنت أرد على الهاتف إذ حين أرفع السماعة كنت أعود لأغلقها مباشرة.

Figure 3: Multi-document Summary Text Sample

NER Text

والاذاعة:O للتليفزيون:O الجديدة:O البرامجية:O الدورة:O اليوم:O تنطلق:O :O الرحبي:I-PERS سالم:B-PERS كتب:O
العديد:O طياتها:O في:O وتحمل:O ويونيو:O ومايو:O ابريل:O اشهر:O طوال:O تستمر:O والتي:O الشباب:O وبرنامج:O
والمستمعين:O المشاهدين:O جميع:O اذواق:O مع:O تتناسب:O التي:O الشيقة:O والفقرات:O الجديدة:O البرامج:O من:O
التنوع:O الاعلام:O وزارة:O في:O المسؤولون:O فيها:O راعى:O الحالية:O البرامج:O دورة:O ..:O سواء:O حد:O على:O
فئاتهم:B-PERS بجميع:O والمستمعين:O المشاهدين:O اوقات:O مراعاة:O الى:O اضافة:O البرامج:O في:O والتجديد:O
من:O منتقاة:O نوعيات:O خلال:O من:O منهجي:O بشكل:O التليفزيون:O لخارطة:O المسبق:O الاعداد:O تم:O حيث:O
بحيث:O الوثائقية:O البرامج:O جدول:O وتغيير:O الاسبوعية:O السهرات:O تشكيلة:O تعديل:O تم:O كما:O البرامج:O
اكدت:O وقد:O ..:O فيها:O التجديد:O المحلية:O البرامج:O طرح:O على:O التركيز:O مع:O الثقافي:O التنوع:O تشمل:O
عام:I-LOC مدير:B-LOC بمكتب:O امس:O ظهر:O عقدته:O الذي:O الصحفي:O اللقاء:O في:O التليفزيون:O ادارة:O
I- محفوظ:B-PERS بن:I-PERS صالح:B-PERS وبوجود:O العبري:I-PERS عبدالله:B-PERS المهندس:O التليفزيون:O
ان:O التليفزيون:O مكتبة:O منسقة:O الراشدي:O وزوينة:O بالتليفزيون:O العامة:O البرامج:O مدير:O القاسمي:PERS
العماني:O المشاهد:O رغبات:O مع:O تتماشى:O متميزة:O بدورة:O الخروج:O اجل:O من:O جاهدة:O سعت:O الادارة:O
الفضائية:O القنوات:O باقي:O بين:O الصحية:O المنافسة:O على:O ايضا:O التركيز:O مع:O الاولى:O بالدرجة:O
في:O تطرحها:O ان:O تريد:O التي:O البرامج:O تختار:O اصبحت:O التليفزيون:O ادارة:O ان:O الى:O مشيرين:O
ارضاء:O اجل:O من:O وذلك:O وجودها:O البرامج:O بعض:O تفرض:O كانت:O ان:O بعد:O البرامجية:O الدورة:O
أمامه:O اللائقة:O بالصورة:O والخروج:O المشاهد:O :O..

Figure 5: Named Entity Recognition Text Sample

Part of Speech Tagged Text

LANCASTER UNIVERSITY

والاذاعة NNS / للتليفزيون DTJJ / الجديدة DTJJ / البرامجية DTNN / الدورة DTNN / اليوم VBP / تنطلق PUNC:/ DTNNP / الرحبي NNP / سالم VBD/
طياتها IN / في NNP / وتحمل NNP / ويونيو NNP / ومايو NN / ابريل NN / اشهر NN / طوال VBP / تستمر NNS / والتي DTNN / الشباب NN / وبرنامج NN
جميع NN / اذواق NN / مع VBP / تتناسب WP / التي DTJJ / الشيقة NNS / والفقرات DTJJ / الجديدة DTNN / البرامج IN / من DTNN / العديد NN
راعى DTJJ / الحالية DTNN / البرامج NN / دورة PUNC./ NN / سواء NN / حد IN / على VN / والمستمعين DTNNS / المشاهدين NOUN_QUANT
مراعاة IN / الى NN / اضافة DTNN / البرامج IN / في NNP / والتجديد DTNN / التنوع DTNN / الاعلام NN / وزارة IN / في DTNNS / المسؤولون NN / فيها
التليفزيون NN / لخارطة DTJJ / المسبق DTNN / الاعداد VBD / تم WRB / حيث NNP / فئاتهم NN / بجميع VN / والمستمعين DTNNS / المشاهدين NN / اوقات
تشكيلة NN / تعديل VBD / تم CC / كما DTNN / البرامج IN / من JJ / منتقاة NNS / نوعيات NN / خلال IN / من JJ / منهجي NN / بشكل DTNNS
مع DTJJ / الثقافي DTNN / التنوع VBP / تشمل NN / بحيث DTJJ / الوثائقية DTNN / البرامج NN / جدول NN / وتغيير DTJJ / الاسبوعية DTNNS / السهرات
ادارة VBD / اكدت NN / وقد PUNC./ NNP / فيها DTNN / التجديد DTJJ / المحلية DTNN / البرامج NN / طرح IN / على DTNN / التركيز NN
التليفزيون NN / عام NN / مدير NN / بمكتب NN / امس NN / ظهر VBD / عقدته WP / الذي DTJJ / الصحفي DTNN / اللقاء IN / في DTNNS / التليفزيون
البرامج NN / مدير DTNNP / القاسمي NNP / محفوظ NNP / بن NNP / صالح NNP / وبوجود DTNNP / العبري NNP / عبدالله DTNN / المهندس DTNNS
سعت DTNN / الادارة IN / ان DTNNS / التليفزيون NN / مكتبة NN / منسقة DTNNP / الراشدي NNP / وزوينة NNP / بالتليفزيون DTJJ / العامة DTNN
العماني DTNN / المشاهد NNS / رغبات NN / مع VBP / تتماشى JJ / متميزة NN / بدورة DTNN / الخروج NN / اجل IN / من NN / جاهدة VBD
القنوات NNS / باقي NN / بين DTJJ / الصحية DTNN / المنافسة IN / على RB / ايضا DTNN / التركيز NN / مع ADJ_NUM / الاولى NNP / بالدرجة DTNNS
ان IN / تريد VBP / التي DTNN / البرامج VBP / تختار VBD / اصبحت DTNNS / التليفزيون NN / ادارة IN / ان IN / الى VN / مشيرين DTJJ / الفضائية
وجودها DTNN / البرامج NOUN_QUANT / بعض VBP / تفرض VBD / كانت IN / ان NN / بعد DTJJ / البرامجية DTNN / الدورة IN / في VBP / تطرحها
NN ./PUNC / أمامه DTJJ / اللائقة NNP / بالصورة NN / والخروج DTNN / المشاهد NN / ارضاء NN / اجل IN / من NN / وذلك NN

Figure 6: Part of Speech Tagger Text Sample

Morphologically Analysed Text

Word	Vowels	Prefix	Stem	Type	Pattern	Root	Suffix
اشهر	اشْهَرْ	#	اشهر	فعل أمر	افْعَلْ	شهر	#
تشكيلة	تَشْكِيلَةِ	#	تشكيلة	مصدر مرة	تَفْعِيلَةِ	شكل	ة: تاء التأنيث
دورة	دَوْرَةُ	#	دورة	مصدر مرة	فَعْلَةُ	دور	ة: تاء التأنيث
طوال	طِوَالُ	#	طوال	مصدر أصلي	فِعَالُ	طول	#
منتقاة	مُنْتَقَاةٍ	#	منتقاة	اسم مفعول	مُفْتَعَاةٍ	نقو	ة: تاء التأنيث
نوعيات	نَوْعِيَّاتِ	#	نوعيات	مصدر صناعي	فَعْلِيَّاتِ	نوع	ات: تاء التأنيث

Figure 4: Morphological Analyser Sample

References:

- Abbas, M., Smaili, K. and Berkani, D. 2011. "Evaluation of Topic Identification Methods on Arabic Corpora". *Journal of Digital Information Management,vol. 9, N. 5, pp.185-192.*
- Benajiba, Y., Rosso, P. and BenedRuiz, J. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. *Computational Linguistics and Intelligent Text Processing, 143–153.*
- El-Haj, M., Kruschwitz, U. and Fox, C. 2010. "Using Mechanical Turk to Create a Corpus of Arabic Summaries". *In The 7th International Language Resources and Evaluation Conference (LREC 2010)., pages 36–39, Valletta, Malta,. LREC.*
- El-Haj, M., Kruschwitz, U. and Fox, C. 2011. "Exploring Clustering for Multi-Document Arabic Summarisation". In The 7th Asian Information Retrieval Societies (AIRS 2011), volume 7097 of Lecture Notes in Computer Science, pages 550–561. Springer Berlin / Heidelberg.
- Koulali, R. and Meziane, A. 2012. "A contribution to Arabic Named Entity Recognition". *In ICT and Knowledge Engineering. ICT Knowledge Engineering, pages 46–52.*
- Mazroui, A., Meziane, A., Ould Abdallahi Ould Bebah, M., Boudlal, A., Lakhouaja, A and Shoul, M. 2011. ALkhalil morphosys: Morphosyntactic analysis system for non voalized Arabic. *In Proceeding of the 7th International Computing Conference in Arabic.*
- Salton G., Wong A. and Yang, S. 2003. "A Vector Space Model for Automatic Indexing". *Proceedings of the Communications of the ACM, 18(11):613–620, 1975.*
- Toutanova, K., Klein, D., Manning, C.D. and Singer, Y. 2003. "Feature-Rich Part-Of-Speech Tagging With a Cyclic Dependency Network". *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03, pages 173–180.*

KALIMAT URL:
https://sourceforge.net/projects/kalimat/

Contact:
Mahmoud El-Haj: **m.el-haj@lancaster.ac.uk**
Rim Koulali: **rim.koulali@gmail.com**