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

1 Introduction

Resources, such as corpora, are important for researchers working on Arabic Natural Language Processing (NLP) (Al-Sulaiti et al. 2006). For this reason we came up with the idea of generating an Arabic multipurpose corpus, which we call KALIMAT[1] (Arabic transliteration of "WORDS"). The automatically created corpus could benefit researchers working on different Arabic NLP areas.

In our work on Arabic we developed, enhanced and tested many Arabic NLP tools. We tuned these tools to provide high quality results. The tools include auto-summarisers, Part of Speech Tagger, Morphological Analyser and Named Entity Recognition (NER). We ran these tools using the same document collection. We provide the output corpus freely for researchers to evaluate their work and to run experiments for different Arabic NLP purposes using one corpus.

2 KALIMAT Corpus

KALIMAT consists of: 1) 20,291 Arabic articles collected from the Omani newspaper Alwatan[2] by (Abbas et al. 2011). 2) 20,291 extractive single-document system summaries. 3) 2,057 multi-document system summaries. 4) 20,291 Named Entity Recognised articles. 5) 20,291 part of speech tagged articles. 6) 20,291 morphologically analyse articles.

The data collection articles fall into six categories: culture, economy, local-news, international-news, religion, and sports. Table 1 shows the collection statistics.

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
Total:	20,291	18,167,183

Table 1: Document Collection Statistics

The reason behind selecting Alwatan's articles was that they contain real text written and used by native Arabic speakers. The collected articles were written by many authors from different backgrounds and they cover a range of topics from different subject areas, each with a credible amount of data. Figure 1 in the Appendix shows a random text sample of Alwatan's articles.

3 Corpus Creation Process

The process of creating KALIMAT was applied to the entire data collection (20,291 articles).

Firstly, we summarised the document collection using two Arabic summarisers, Gen–Summ and Arabic Cluster-based.

Gen-Summ (El-Haj et al. 2010) is a single document summariser based on the VSM model (Salton et al. 1975) that takes an Arabic document and its first sentence and returns an extractive summary. A number of 20,291 system summaries have been generated. 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 extractive system 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. Table 2 shows the multi-document summaries distribution. Figures 2 and 3 show samples of the generated single and multi document summaries.

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

Table 2: Multi-document Summaries Statistics

Secondly, 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. To annotate the data collection we followed the Computational Natural Language Learning (CoNLL) 2002[3] and 2003[4]

[1] http://www.lancs.ac.uk/staff/elhaj/corpora.htm
[2] http://www.alwatan.com/
[3] http://www.cnts.ua.ac.be/conll2002/
[4] http://www.cnts.ua.ac.be/conll2003/

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 and include date, time, number, monetary expressions, measurement expressions and percentages.

ANER system was trained using ANERCorpus (Benajiba et al. 2007), a manually annotated corpus following the CoNLL shared task. The reason behind choosing ANERCorpus to train our system was that the corpus articles were chosen from Arabic newswires and Wikipedia Arabic, which is quite close to Alwatan's data collection, see Section 2.

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.

To improve the performance of the developed ANER system, an automatic pattern extractor framework was implemented. The extracter was based on POS tags information and linguistic filters including trigger words and stop-word elimination. The ANER system achieved an overall F-measure of 83.20%.

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. Figure 5 in the Appendix shows a text sample annotated using ANER.

Thirdly, we used Stanford POSTagger (Toutanova et al. 2003) to annotate the 20,291 document collection. The system is a Java implementation of the log-linear part-of-speech taggers. 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 modeling of unknown word features

The Stanford POSTagger is a supervised system depending on different trained models for many languages including Arabic. The accurate model for Arabic was trained using the Arabic Tree-bank p1-3 corpus based on maximum entropy and using augmented Bies 5 mapping of ATB tags. 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. Figure 6 in the Appendix shows the output of the Stanford POSTagger.

Finally, we applied a morphological analysis process on the data collection using Alkhalil morphological analyser (Mazroui et al. 2011).

Alkhalil[6] Morphological Analyzer 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 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), see Figure 4 in the Appendix.

Applying Alkhalil analyser on the data collection we reached an accuracy of 96%. We implemented a Viterbi algorithm to get one solution that is relevant to the context of the analyzed article.

The morphological analysis of 20,291 documents could help in improving the performance of many tools such as: automatic vocalization, spell checking, automatic summarization. See Figure 4 for a morphologically analysed text sample using Alkhalil's system.

[5] http://www.ircs.upenn.edu/arabic/Jan03release/arabic-POStags-collapse-to-PennPOStags.txt

[6] http://sourceforge.net/projects/alkhalil/

4 Conclusion

We provide KALIMAT[7] for free including the articles, annotated text, entities and summaries to help advancing the work on Arabic NLP. The corpus can be downloaded directly from: https://sourceforge.net/projects/kalimat/.

The corpus and the results we achieved can be used by researchers as gold-standards and or baselines to test and evaluate their Arabic tools. We also welcome any amendments to the corpus by other researchers.

In our work we address the shortage of relevant data for Arabic natural language processing, taking into consideration the lack of Arab participants to come up with resources that are important for researchers working on Arabic NLP.

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.

Al-Sulaiti, L., Atwell, ES. and Steven, E. 2006. "The design of a corpus of Contemporary Arabic". *International Journal of Corpus Linguistics*, 11(2): 135–171.

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.

[7] https://sourceforge.net/projects/kalimat/

Appendix

KALIMAT corpus consists of 20,291 news articles collected from Alwatan newspaper (Abbas et al.). From the original HTML text only unformatted content text was kept, removing any images, tables or hyperlinks. We applied different NLP tools on the data collection to generate 20,291 single-document summaries, 2,057 multi-document summaries, and 20,291 Named Entity Recognition annotated articles. In addition to 20,291 part-of-speech tagged articles and 20,291 morphologically analysed articles. Figures 1 to 6 show samples of KALIMAT annotated text.

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

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

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

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	#	نوعيا ت	مصدر صناعي	فَعْلِيَّاتِ	نوع	ات:تاء التأنيث
وزارة	وِزَارَةُ	#	وزارة	اسم جامد	فِعَالَةُ	وزر	ة: تاء التأنيث

Figure 4: Morphological Analyser Sample

:O كتب:B-PERS سالم:I-PERS الرحبي:O :O: تنطلق:O اليوم:O الدورة:O البرامجية
:Oالجديدة:O للتليفزيون:O والاذاعة:O وبرنامج:O الشباب:O والتي:O تستمر:O طوال
:Oاشهر:O ابريل:O ومايو:O ويونيو:O وتحمل:O في:O طياتها:O العديد:O من:O البرامج
:Oالجديدة:O والفقرات:O الشيقة:O التي:O تتناسب:O مع:O اذواق:O جميع:O المشاهدين
:Oوالمستمعين:O على:O حد:O سواء:O .:O دورة:O البرامج:O الحالية:O راعى:O فيها
:Oالمسؤولون:O في:O وزارة:O الاعلام:O التنوع:O والتجديد:O في:O البرامج:O اضافة
:Oالى:O مراعاة:O اوقات:O المشاهدين:O والمستمعين:O بجميع:B-PERS فئاتهم:O حيث
:Oتم:O الاعداد:O المسبق:O لخارطة:O التليفزيون:O بشكل:O منهجي:O من:O خلال
:Oنوعيات:O منتقاة:O من:O البرامج:O كما:O تم:O تعديل:O تشكيلة:O السهرات
:Oالاسبوعية:O وتغيير:O جدول:O البرامج:O الوثائقية:O بحيث:O تشمل:O التنوع
:Oالثقافي:O مع:O التركيز:O على:O طرح:O البرامج:O المحلية:O التجديد:O فيها .:O
:Oوقد:O اكدت:O ادارة:O التليفزيون:O في:O اللقاء:O الصحفي:O الذي:O عقدته:O ظهر
:Oامس:O بمكتب:B-LOC مدير:I-LOC عام:O التليفزيون:O المهندس:B-PERS عبدالله
:I-PERSالعبري:O وبوجود:B-PERS صالح:I-PERS بن:B-PERS محفوظ I-
:PERSالقاسمي:O مدير:O البرامج:O العامة:O بالتليفزيون:O وزوينة:O الراشدي
:Oمنسقة:O مكتبة:O التليفزيون:O ان:O الادارة:O سعت:O جاهدة:O من:O اجل
:Oالخروج:O بدورة:O متميزة:O تتماشى:O مع:O رغبات:O المشاهد:O العماني
:Oبالدرجة:O الاولى:O مع:O التركيز:O ايضا:O على:O المنافسة:O الصحية:O بين
:Oباقي:O القنوات:O الفضائية:O مشيرين:O الى:O ان:O ادارة:O التليفزيون:O اصبحت
:Oتختار:O البرامج:O التي:O تريد:O ان:O تطرحها:O في:O الدورة:O البرامجية:O بعد
:Oان:O كانت:O تفرض:O بعض:O البرامج:O وجودها:O وذلك:O من:O اجل:O ارضاء
:Oالمشاهد:O والخروج:O بالصورة:O اللائقة:O أمامهO :.

Figure 5: Named Entity Recognition Text Sample

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

Figure 6: Part of Speech Tagger Text Sample