University of Essex at the TAC 2011 Multilingual Summarisation Pilot

Mahmoud El-Haj, Udo Kruschwitz, Chris Fox
School of Computer Science and Electronic Engineering
University of Essex
{melhaj,udo,foxcj}@essex.ac.uk

Abstract

We present the results of our Arabic and English runs at the TAC 2011 Multilingual summarisation (MultiLing) task. We participated with centroid-based clustering for multi-document summarisation. The automatically generated Arabic and English summaries were evaluated by human participants and by two automatic evaluation metrics, ROUGE and AutoSummENG. The results are compared with the other systems that participated in the same track on both Arabic and English languages. Our Arabic summariser performed particularly well in the human evaluation.

1 Introduction

Multi-document summarisation is the process of producing a single summary of a collection of related documents. TAC 2011 announced a new task this year in which for the first time participants were able to run their summarisers on different languages having a corpus and gold standard summaries in seven different languages, namely Arabic, Czech, English, French, Greek, Hebrew and Hindi. The task is called Multilingual Summarization pilot (MultiLing), which aims to foster and promote the use of multi-lingual algorithms for summarization. This includes the effort of transforming an algorithm or a set of resources from a mono-lingual to a multi-lingual version. Participating summarisers were expected to be language-independent and each participant was required to submit results for at least two of the seven languages (in our case Arabic and English). The MultiLing task requires the participant to generate a single, fluent, representative summary from a set of documents describing an event sequence. The output summary is expected to be between 240 and 250 words[1] (inclusive).

We show the results of our participation at the TAC 2011 MultiLing pilot where we applied partitional clustering technique on our multi-document summarisers. We compare our results against a range of different systems that participated in this task.

The paper is structured as follows. We will start with a discussion of related work in Section 2. We will give a brief summary about the dataset and evaluation metrics used in Section 3. Section 4 will describe the clustering approaches we are working with together with a description of our summarisers and the experimental setup. Results are discussed in Section 5, and we conclude in Section 6.

2 Related Work

2.1 Multi-Document Summarisation

The analysis for multi-document summarisation is usually performed at the level of sentences or documents. Multi-document summarisation systems follow two approaches: *extractive* or *abstractive*. Work on abstractive summarisation is limited; it requires natural language analysis and generation techniques that are still not that robust. For this reason, most of the current summarisation systems rely on the extractive approach.

One early multi-document summariser used information extraction (IE) to identify similarities and dif-

[1] The count of words was provided by the *wc -w* linux command.

ferences between documents (McKeown and Radev, 1995). Later systems combined IE with a process that regenerates the extracted units in order to improve the quality of the summarisation (A. Funk and Bontcheva, 2007). (Zhao et al., 2009) describe a method for query-focused multi-document summarisation.

In our own work we do not focus on query-based summarisation as we are focusing on query-independent summarisation, which is the main objective of the MultiLing pilot.

Summarisation of Arabic documents has not advanced as fast as work in other languages such as English. The summariser "Lakhas" (Douzidia and Lapalme, 2004) was developed using extraction techniques to produce ten-words summaries of news articles. (Turchi et al., 2010) presented a method for evaluating multilingual, multi-document, extractive summarisation, using a parallel corpus of seven languages. In their approach, the most important sentences in a document collection were manually selected in one language. This gold-standard summary was then projected into the other languages in the parallel corpus.

We have previously applied centroid-based clustering to summarise multiple documents in Arabic and English (El-Haj et al., 2011a) and (El-Haj et al., 2011b). Here we wanted to find out how these language-independent summarisers perform against other approaches, and we were also interested in finding out how each of the two languages performed.

2.2 Clustering for Summarisation

Data clustering is the assignment of a set of observations into subsets, so called clusters. As a method of unsupervised learning, clustering has received a lot of attention in past years to improve information retrieval (IR) or to enhance the quality of multi-document summarisation (Dunlavy et al., 2007). Clustering has been applied to many document levels starting from the document itself down to sentences and words. Clustering can broadly be grouped into hierarchical clustering and partitional clustering.

In centroid-based multi-document summarisation, a form of partitional clustering, similarity to the cluster centroid and the top ranked sentences has been the main factor in clustering sentences, where the centroid is defined as a pseudo-document consisting of words with TF*IDF scores greater than a predefined threshold (Radev et al., 2000) and (Radev et al., 2004). We will also focus on centroid-based clustering.

The work by (Liu and Lindroos, 2006) proposed a Chinese multi-document summariser which is based on clustering paragraphs of the input articles. In their work they cluster the source documents by paragraph units instead of sentences and the cluster size used changes based on the number of extracted sentences.

(Wan and Yang, 2008), proposed a multi-document summarisation technique using cluster-based link analysis, in their work they used three clustering detection algorithms including k-means, agglomerative and divisive clustering. The proposed module relies on clustering the sentences into different themes (subtopics), the number of clusters is defined by taking the absolute square root of the number of all sentences in the document set. We use a predefined range for the number of clusters.

(Sarkar, 2009), presented a sentence clustering based multi-document summarisation system by adopting the incremental clustering method which has been used for web clustering in (Hammouda and Kamel, 2004). In their work they reordered the clusters based on their sizes measured in terms of sentence-counts assuming the more sentences in a cluster the more important the cluster is.

To the best of our knowledge, little work has been reported on applying clustering for Arabic multi-document summarisation. (Schlesinger et al., 2008) presented CLASSY, an Arabic/English query-based multi-document summariser system. They used an unsupervised modified k-means method to iteratively cluster multiple documents into different topics (stories). They relied on the automatic translation of an Arabic corpus into English. At the time of their experiments, the quality of machine translation was not high enough. This led to difficulties in reading and understanding the translated dataset. The translation resulted in inconsistent sentences; core keywords may have been dropped when translating. Errors in tokenisation and sentence-splitting were among the main challenges.

3 Dataset and Evaluation Metrics

3.1 Test Collection

The test collection for the MultiLing pilot is available in the previously mentioned 7 languages[2]. The dataset is based on WikiNews texts[3]. The source documents contain no meta-data or tags and are represented as UTF–8 plain text files. The dataset of each language contains 100 articles divided into 10 reference sets, each contains 10 related articles discussing the same topic. The original language of the dataset is English. The organisers of the pilot were responsible for translating the corpus into different languages by having native speaker participants for each of the 7 languages. In addition to the news articles the dataset also provides human-generated multi-document gold standard summaries.

3.2 Evaluation

Evaluating the quality and consistency of a generated summary has proven to be a difficult problem (Fiszman et al., 2009). This is mainly because there is no obvious ideal, objective summary. Two classes of metrics have been developed: form metrics and content metrics. Form metrics focus on grammaticality, overall text coherence, and organisation. They are usually measured on a point scale (Brandow et al., 1995). Content metrics are more difficult to measure. Typically, system output is compared sentence by sentence or unit by unit to one or more human-generated ideal summaries. As with information retrieval, the percentage of information presented in the system's summary (precision) and the percentage of important information omitted from the summary (recall) can be assessed.

There are various models for system evaluation that may help in solving this problem. Automatic evaluation metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and AutoSummENG (Giannakopoulos et al., 2008) have been shown to correlate well with human evaluations for content match in text summarisation and machine translation. Other commonly used evaluations include assessing readers' understanding of automatically generated summaries. Human-performed evaluation may be preferable to automatic methods, but the cost is high.

For this task, the evaluation of results was performed both automatically and manually. The manual evaluation was based on the *Overall Responsiveness* of a text and the automatic evaluation used the *ROUGE* and *AutoSummENG-MeMoG* methods to provide a grading of performance. For the manual evaluation the human evaluators were provided with the following guidelines: Each summary is to be assigned an integer grade from 1 to 5, related to the overall responsiveness of the summary. We consider a text to be worth a 5, if it appears to cover all the important aspects of the corresponding document set using fluent, readable language. A text should be assigned a 1, if it is either unreadable, nonsensical, or contains only trivial information from the document set. We consider the content and the quality of the language to be equally important in the grading.

Summaries that are out-of-limit are penalised using the Length-Aware Grading measure (LAG). Given a summary S of length $|S|$ (in words) assigned a grade g, a lower word limit count l_{min} and an upper word limit count l_{max}, then LAG is defined as follows.

$$LAG(g,S) = g{*}(1 - \frac{\max(\max(l_{min} - |S|, |S| - l_{max}), 0)}{l_{min}}) \tag{1}$$

In the task specific evaluation, $l_{min} = 240, l_{max} = 250$. LAG simply provides a linearly diminishing weight to grades diverging from the limits.

The automatic evaluation was based on human-generated model summaries provided by fluent speakers of each corresponding language (native speakers in the general case). The models used were, ROUGE variations (ROUGE1, ROUGE2, ROUGE-SU4) (Lin, 2004) and the MeMoG variation (Giannakopoulos and Karkaletsis, 2010) of AutoSummENG (Giannakopoulos et al., 2008).

4 Cluster-based Summarisation

For all our experiments we are using a generic multi-document extractive summariser that has been implemented for both Arabic and English (using identical processing pipelines for both languages). Summaries are generated by selecting sentences from sets of related articles. Details on the summarisers used in this task are given in (El-Haj et al., 2011a) and (El-Haj et al., 2011b).

[2] http://www.nist.gov/tac/2011/Summarization/index.html

[3] http://en.wikinews.org/wiki/Main_Page

We will now describe the clustering methods employed in our experiments, the actual summarisation process and the experimental setup.

4.1 K-means Clustering

K-means clustering is a partitional centroid-based clustering algorithm. The algorithm randomly selects a number of sentences as the initial centroids, the number of sentences is dependent on the cluster size assigned. The algorithm then iteratively assigns all sentences to the closest cluster, and recalculates the centroid of each cluster, until the centroids no longer change. For our experiments, the similarity between a sentence and a cluster centroid is calculated using the standard cosine measure applied to tokens within the sentence.

4.2 Centroid-Based Clustering Experiment

In previous work we experimented with partitional clustering summarisation (El-Haj et al., 2011a). We found that clustering sentences and then selecting sentences from the biggest cluster performed very well. We experimented with different numbers of clusters and found the extreme case of a single cluster was among the best-performing settings. Based on the findings we decided to enter TAC 2011 MultiLing with a clustering approach that treats all documents to be summarised as a single bag of sentences, and the sentences are clustered using a *single cluster*. We can then rank all sentences in order of similarity to the centroid. The summary is generated by selecting sentences in that ranked order, i.e. selecting sentences that are close to the centroid, until the expected limit is reached. In the resulting summary we keep the order of sentences as they appear in the clusters (i.e. a sentence very similar to the centroid appears earlier on in the summary than one that is less similar).

The intuition for this approach is the assumption that a single cluster will give a coherent summary all centred around a single theme, where other approaches expect to result in summaries that contain more aspects of the topics discussed in the documents and therefore a summary that gives a broader picture.

Note that in our experiments we trim the resulting summary to a particular length. As indicated in the task, the acceptable limits for the word count of a summary were between 240 and 250 words (inclusive).

SysID	Human (Overall)	Human (LAG)
ID1	3.77	3.77
ID9	3.73	3.73
ID8	3.70	3.66
ID3	3.43	3.30
ID7	3.30	3.20
ID10	3.20	3.10
ID2	3.10	3.10
ID6	3.10	2.76
ID4	2.77	2.76

Table 1: Arabic Overall and LAG Responsiveness Scores

SysID	Human (Overall)	Human (LAG)
ID3	3.83	3.55
ID2	3.53	3.53
ID10	3.20	3.20
ID1	3.20	3.10
ID5	3.03	2.92
ID8	2.73	2.73
ID9	2.50	2.50
ID7	2.30	2.29
ID6	2.67	2.20
ID4	2.033	2.033

Table 2: English Overall and LAG Responsiveness Scores

5 Results and Discussion

Apart from the actual participants in the MultiLing task there were also a global *baseline* (System ID9) and a global *topline* (System ID10).

For the Arabic language, there were 7 participants (peers) in addition to the two baseline systems, for a total of 9 runs. The English language had 8 participants in addition to the two baseline systems, for a total of 10 runs.

Our system in both the Arabic and the English competition is referred to as **ID8**.

Tables 1 and 2 illustrate the overall and the length-aware grading measure (LAG) for systems participating for the Arabic and English language respectively. As we can see in the first table, our system (ID8) scored very well, only two systems (one of them the baseline) scored better but not by a big margin. The LAG grade of our system reflect that some of our summaries were out of limit (below 240 words), but as we can see this did not affect the ranking.

SysID	Recall	Precision	F-score
ID10	0.46751	0.25828	0.30786
ID3	0.37218	0.29644	0.29987
ID2	0.34194	0.29444	0.29188
ID6	0.35648	0.25396	0.2763
ID8	0.38854	0.22008	0.26786
ID4	0.42259	0.20676	0.26279
ID1	0.29869	0.21359	0.2319
ID9	0.32405	0.23596	0.23097
ID7	0.24058	0.22703	0.22376

Table 3: Arabic ROUGE-1 Scores

SysID	Recall	Precision	F-score
ID10	0.23394	0.13669	0.14922
ID3	0.15808	0.13857	0.1278
ID6	0.13767	0.0992	0.10629
ID2	0.13858	0.09774	0.10347
ID8	0.14726	0.07851	0.09653
ID9	0.12559	0.10451	0.09497
ID1	0.12057	0.08482	0.0889
ID4	0.13962	0.06886	0.08634
ID7	0.10627	0.07772	0.08577

Table 4: Arabic ROUGE-2 Scores

Observing the results of the English human evaluation in Table 2, we see that our system performed better than the baseline. However, we note that the scores given by human assessors is substantially lower than those for the Arabic system.

Tables 3, 4 and 5 illustrate the ROUGE results and the ranking of our Arabic multi-document summariser (System ID8) as well as the corresponding AutoSummENG-MeMoG evaluation metric in Table 9. The ROUGE results correlate quite closely with the AutoSummENG–MeMoG ranking of systems.

We observe that the automatic evaluation results place our Arabic summariser further down in the ranked lists of systems compared to the human assessment.

Tables 6, 7 and 8 give the ROUGE results and the ranking of our English multi-document summariser, Table 10 has the AutoSummENG-MeMoG evaluation results. Like with the Arabic summariser, we note that the human assessment of our run places our system higher in the ranked order than the automatic evaluation scores.

SysID	Recall	Precision	F-score
ID10	0.2783	0.14152	0.15489
ID3	0.19889	0.16758	0.1514
ID2	0.16618	0.14293	0.13309
ID6	0.17617	0.1145	0.12456
ID8	0.18475	0.09219	0.11487
ID4	0.20836	0.07856	0.1071
ID7	0.11818	0.09413	0.09874
ID1	0.14033	0.09419	0.09871
ID9	0.15185	0.11618	0.0974

Table 5: Arabic ROUGE-SU4 Scores

SysID	Recall	Precision	F-score
ID10	0.52488	0.51806	0.52141
ID2	0.46481	0.45655	0.46062
ID3	0.43169	0.47909	0.45404
ID4	0.44423	0.44966	0.44691
ID5	0.41092	0.43513	0.42243
ID1	0.40524	0.41253	0.40776
ID6	0.3547	0.45122	0.39617
ID7	0.39586	0.3953	0.39547
ID8	0.38714	0.39265	0.38985
ID9	0.38105	0.37726	0.3791

Table 6: English ROUGE-1 Scores

SysID	Recall	Precision	F-score
ID10	0.25177	0.2483	0.25
ID3	0.1733	0.19256	0.18237
ID2	0.17052	0.16779	0.16914
ID4	0.1517	0.15369	0.15269
ID5	0.13605	0.14404	0.13985
ID1	0.12125	0.12448	0.12247
ID8	0.12144	0.12298	0.12219
ID6	0.10655	0.1367	0.11937
ID9	0.10962	0.10841	0.109
ID7	0.09662	0.09612	0.09635

Table 7: English ROUGE-2 Scores

SysID	Recall	Precision	F-score
ID10	0.27248	0.26882	0.27062
ID1	0.15995	0.16322	0.16112
ID2	0.2022	0.19868	0.20042
ID3	0.19927	0.22148	0.20973
ID4	0.19083	0.1932	0.192
ID5	0.17475	0.18503	0.17964
ID6	0.1457	0.18648	0.16312
ID7	0.14507	0.1446	0.1448
ID8	0.1566	0.15874	0.15765
ID9	0.14805	0.14655	0.14728

Table 8: English ROUGE-SU4 Scores

SysID	MeMoG
ID10	0.665674
ID3	0.482755
ID4	0.382946
ID2	0.368587
ID6	0.340396
ID8	0.305233
ID1	0.296868
ID9	0.282094
ID7	0.261209

Table 9: Arabic AutoSummENG–MeMoG Scores

SysID	MeMoG
ID10	0.5477871
ID3	0.4256148
ID2	0.3859586
ID4	0.3785725
ID5	0.3500278
ID6	0.3490875
ID1	0.3443412
ID8	0.3323676
ID7	0.3108508
ID9	0.304319

Table 10: English AutoSummENG–MeMoG Scores

6 Conclusion

In this paper we presented the results of our participation in the TAC 2011 MultiLing pilot. We submitted results for multi-document summarisation systems in two languages, Arabic and English. We applied a simple clustering approach for multi-document Arabic and English summarisation where the summary consists of a set of sentences selected from all documents that are most similar to the centroid (sentence) of the entire document set.

Based on human assessments, we found that our approach appears to work very well for Arabic but less so for English. We also found that the automatic evaluation scores rank both our system further down the ranked list of submissions than the human assessment scores.

There is a lot of room for further investigation. One interesting point would be to see which of the differences are actually statistically significant and how much variation there is between the individual scores.

Among our own future work is the application of more fine-tuned clustering to improve the results.

References

[A. Funk and Bontcheva2007] H. Saggion A. Funk, D. Maynard and K. Bontcheva. 2007. Ontological integration of information extracted from multiple sources. In *In the Multi-source Multilingual Information Extraction and Summarization (MMIES) workshop at Recent Advances in Natural Language Processing (RANLP07)*, Borovets, Bulgaria. RANLP07.

[Brandow et al.1995] Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Inf. Process. Manage.*, 31:675–685, September.

[Douzidia and Lapalme2004] F. S. Douzidia and G. Lapalme. 2004. Lakhas, an arabic summarising system. In *In the Proceedings of the Document Understanding Conferences (DUC) Workshop*, pages 128–135. DUC.

[Dunlavy et al.2007] Daniel M. Dunlavy, Dianne P. O'Leary, John M. Conroy, and Judith D. Schlesinger. 2007. Qcs: A system for querying, clustering and summarizing documents. *Inf. Process. Manage.*, 43:1588–1605, November.

[El-Haj et al.2011a] M. El-Haj, U. Kruschwitz, and C. J. Fox. 2011a. Exploring clustering for multi-document arabic summarisation. In *The Seventh Asia Information*

Retrieval Societies Conference (AIRS 2011), Dubai, UAE. Springer LNCS.

[El-Haj et al.2011b] M. El-Haj, U. Kruschwitz, and C. J. Fox. 2011b. Multi-document arabic text summarisation. In *3rd Computer science and Electronic Engineering Conference (CEEC'11)*, pages 40–44, Colchester, UK. IEEE Computer Society.

[Fiszman et al.2009] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T. C. Rindflesch. 2009. Automatic summarization of medline citations for evidence-based medical treatment: A topic-oriented evaluation. *Jouranl of Biomedical Informatics*, 42(5):801–813.

[Giannakopoulos and Karkaletsis2010] George Giannakopoulos and Vangelis Karkaletsis. 2010. Summarization system evaluation variations based on n-gram graphs. In *Proceedings of Text Analysis Conference 2010*. NIST, Gaithersburg, MD, USA.

[Giannakopoulos et al.2008] G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. 2008. Summarization system evaluation revisited: N-gram graphs. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(3):1–39.

[Hammouda and Kamel2004] Khaled M. Hammouda and Mohamed S. Kamel. 2004. Efficient phrase-based document indexing for web document clustering. *IEEE Trans. on Knowl. and Data Eng.*, 16:1279–1296, October.

[Lin2004] C. Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, pages 25–26.

[Liu and Lindroos2006] S. Liu and J. Lindroos. 2006. Towards fast digestion of imf staff reports with automated text summarization systems. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 978–982. IEEE Computer Society.

[McKeown and Radev1995] Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 74–82, New York, NY, USA. ACM.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Radev et al.2000] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 21–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Radev et al.2004] Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938.

[Sarkar2009] Kamal Sarkar. 2009. Centroid-based summarization of multiple documents. *TECHNIA - International Journal of Computing Science and Communication Technologies*, 2.

[Schlesinger et al.2008] Judith D. Schlesinger, Dianne P. O'Leary, and John M. Conroy. 2008. Arabic/english multi-document summarization with classy: the past and the future. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, CICLing'08, pages 568–581, Berlin, Heidelberg. Springer-Verlag.

[Turchi et al.2010] Marco Turchi, Josef Steinberger, Mijail Kabadjov, and Ralf Steinberger. 2010. Using parallel corpora for multilingual (multi-document) summarisation evaluation. In *Proceedings of the 2010 international conference on Multilingual and multimodal information access evaluation: cross-language evaluation forum*, CLEF'10, pages 52–63, Berlin, Heidelberg. Springer-Verlag.

[Wan and Yang2008] Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 299–306, New York, NY, USA. ACM.

[Zhao et al.2009] Lin Zhao, Lide Wu, and Xuanjing Huang. 2009. Using query expansion in graph-based approach for query-focused multi-document summarization. *Inf. Process. Manage.*, 45:35–41, January.