Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Creating Language Resources for Under-resourced Languages: Methodologies, and experiments with Arabic

Mahmoud El-Haj · **Udo Kruschwitz** · **Chris Fox**

Received: 08/11/2013 / Accepted: 15/07/2014

Abstract Language resources are important for those working on computational methods to analyse and study languages. These resources are needed to help advancing the research in fields such as natural language processing, machine learning, information retrieval and text analysis in general. We describe the creation of useful resources for languages that currently lack them, taking resources for Arabic summarisation as a case study. We illustrate three different paradigms for creating language resources, namely: (1) using crowdsourcing to produce a small resource rapidly and relatively cheaply; (2) translating an existing gold-standard dataset, which is relatively easy but potentially of lower quality; and (3) using manual effort with appropriately skilled human participants to create a resource that is more expensive but of high quality. The last of these was used as a test collection for TAC-2011. An evaluation of the resources is also presented.

The current paper describes and extends the resource creation activities and evaluations that underpinned experiments and findings that have previously appeared as an LREC workshop paper (El-Haj et al 2010), a student conference paper (El-Haj et al 2011b), and a description of a multilingual summarisation pilot (El-Haj et al 2011c; Giannakopoulos et al 2011).

M. El-Haj
School of Computing and Communications, Lancaster University, UK
Tel.: +44(0)1524 51 0348
E-mail: m.el-haj@lancaster.ac.uk

U. Kruschwitz
CSEE, University of Essex, UK
Tel.: +44 (0)1206 87 2669
E-mail: udo@essex.ac.uk

C. Fox
CSEE, University of Essex, UK
Tel.: +44 (0)1206 87 2576
E-mail: foxcj@essex.ac.uk

1 Introduction

1.1 Motivation

Language resources are important for researchers working on using computational methods to analyse and study languages. These resources are needed to help make advances in natural language processing (NLP), machine learning, information retrieval, and text analysis in general (see for example Jing and McKeown 1998; Wilks et al 1988; Schalley 2012; Boyer and Brun 2007; Halpern 2006; de Chalendar and Nouvel 2009; El-Haj et al 2010; Walther and Sagot 2010).

Researchers have developed a wide range of NLP tools to analyse, parse and annotate different languages automatically. Language resources play two roles in these activities. The first is the use of large-scale annotated corpora to drive statistical NLP techniques. The second is the need for test collections for the purpose of evaluation against a gold-standard. Such resources for NLP are documented by efforts such as the Language Resources and Evaluation (LRE) Map (Calzolari et al 2010). But for some languages, there are few such resources. In our case, the lack of gold-standard resources for the problem of summarising documents in Arabic lead us to investigate methods for developing them. Our investigations should be relevant to other languages and applications.[1]

Arabic is an appropriate example to consider. Despite being a widely spoken language, it has been widely acknowledged that it has few publicly available tools and resources, apart from a few notable exceptions, such as the Arabic Penn Treebank[2] (Maamouri et al 2003, 2004, 2005; Mourad and Darwish 2013), the Prague Arabic Dependency Treebank[3] (Hajic et al 2004), and document collections such as the Arabic Gigaword corpus (Graff 2003; Graff et al 2006). In particular, Arabic NLP lacks resources such as corpora, lexicons, machine-readable dictionaries in addition to fully automated fundamental NLP tools such as tokenizers, part-of-speech taggers, parsers, and semantic role labelers (Diab et al 2007), although this has started to change in recent years (Maegaard et al 2008; Alghamdi et al 2009; Habash and Roth 2011; Diehl et al 2012).[4] Benajiba et al (2010) built an Arabic Named Entity system by bootstrapping noisy features based on a projection from a parallel Arabic–English corpus.

As is the case with many other NLP applications, most of the activities in our problem domain of text summarisation are concerned with the English language. This is evident in the long track record of competitions as part of the Text Analysis Confer-

[1] Other language cited in the literature as suffering from a lack of resources include Sorani Kurdish (Walther and Sagot 2010) and Swahili (Nganga 2012). As can be seen from the META-NET whitepaper series (http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison) some European languages also suffer from weak, or no support.

[2] http://www.ircs.upenn.edu/arabic/

[3] http://ufal.mff.cuni.cz/padt/PADT_1.0/

[4] Some contributing reasons for the shortage of resources and tools for Arabic may in part be due to its complex morphology, the absence of diacritics (vowels) in written text and the fact that Arabic does not use capitalisation, which causes problems for named entity recognition (Benajiba et al 2009).

ence series (TAC)[5] and, prior to that, the Document Understanding Conference series (DUC)[6], which have both focused on English. This is reflected in the availability of resources: in particular, at the time we started our research into Arabic summarisation there were no readily available Arabic gold-standards for evaluating the results. We needed to consider how to generate such resources, while reducing the cost and time taken compared to traditional approaches.

1.2 Overview

We will demonstrate three methods for acquiring gold-standard language resources for Arabic summarisation:

Crowdsourcing: using *crowdsourcing* to acquire a resource that is relatively cheap and small but can be built rapidly;

Machine translation: using existing gold-standard collections and *machine translation* tools to translate an existing gold-standard resource that is easy to create but of lower quality, with more "noise";

Manual methods: using manual effort with *native speakers* to create a resource that is more expensive but of high quality.

These demonstrate how additional resources can be created rapidly and relatively cheaply where currently they are in short supply. The resources themselves should also be of benefit to those working on Arabic NLP. As an example, the specific resources created using the last of these methods was used as a reference collection in the TAC-2011 MultiLing workshop[7], illustrating how such resources can contribute to advancing the state-of-the-art.

The paper is organised as follows. Section 2 presents the background of the different paradigms and methodologies for creating language resources, and some background on Arabic and NLP tools for Arabic. Section 3 illustrates three methods for creating language resources, using Arabic gold-standard summarisation as a case study. In Section 4 we present an evaluation of the resources we created. Section 5 concludes.

2 Related Work

We briefly review three areas of related work: language resource creation; text summarisation; and Arabic natural language processing.

2.1 Creating Language Resources

Language resources play an important role in the development and evaluation of NLP tools and techniques. The creation of such resources can be performed either manu-

[5] http://www.nist.gov/tac/
[6] http://duc.nist.gov/
[7] http://www.nist.gov/tac/2011/Summarization/

ally or automatically. Manual creation of resources usually involves human experts who may, for example, annotate or translate such datasets. In the case of automatic summarisation, the experts would create gold-standard summaries for a dataset to assess the quality of the automated process. Automatic creation of language resources makes essential use of computers where we might otherwise have employed an expert. In some cases a combination of automatic and manual techniques is used (as with the English Penn Treebank, Marcus et al 1993). Marsi and Krahmer (2013) created a paraphrase annotated corpus in Dutch. The corpus comprises over 2.1M tokens, 678K of which was manually annotated and 1,511K was automatically processed.

2.1.1 Manual Creation of Language Resources

Traditionally, language resources such as the lexical knowledge base *WordNet* (Fellbaum 1998; Abouenour et al 2013), were created manually by experts. A significant amount of time and effort is required. The results can be of *high quality*, but they may be somewhat narrow in scope. The manual approach is still popular, even if some details began to change around twenty years ago, when large volumes of textual data became readily available in electronic form. This has allowed statistical "bootstrapped" approaches to be adopted.

A key example of expert annotation is the English Penn Treebank annotated corpus (Marcus et al 1993). Although this used an automatic phase for some of the initial annotations, this was followed by manual correction. Manually created corpora have been useful in developing statistical NLP methods for a range of problems. Examples include a corpus for detecting plagiarism (zu Meyer et al 2007), for paraphrasing (Dolan et al 2004), and for machine translation (Huang et al 2002, for example).

Outahajala et al (2011) built a POS-tagger for Amazighe, an under-resourced language. The data used to accomplish the work was manually collected and annotated. To help increasing the performance of the tagger, they used machine learning (SVM) and other resources, such as dictionaries and word segmentation tools to process the text and extract feature set consisting of lexical context and character n-grams. The corpus contained 20,000 tokens and was used to train their POS-tagger model.

More recently *crowdsourcing* approaches have become popular. While still manual in character, the role of the expert is replaced by the votes of a number of relatively naive users. This can help reduce the effort and expense of creating large-scale language resources with no significant adverse impact on annotation quality, at least for some judgements (see for example Howe 2008; Snow et al 2008). Typically, crowdsourcing involves soliciting contributions from a community of online users. One of the main platforms applied to a variety of natural language processing tasks in this context is Amazon's Mechanical Turk (AMT).[8]

Snow et al (2008) examined the accuracy of applying AMT for different NLP tasks, namely affect recognition, word similarity, recognising textual entailment, event temporal ordering, and word sense disambiguation. Annotations by AMT non-expert workers showed high agreement with gold-standard labels provided by expert annotators. Similar results concerning quality have been demonstrated by other studies,

[8] http://www.mturk.com/

such as in the context of translation (Callison-Burch 2009).[9] Using crowdsourcing to create language resources has emerged as an economical and fast alternative to human experts, but care must be taken to guarantee high-quality results. Different factors may affect the overall quality obtained from the annotation (Aker et al 2012).

Examples of crowdsourcing approaches as a useful method for creating natural language processing resources include Kaisser and Lowe (2008); Yang et al (2009); Alonso and Mizzaro (2009); Callison-Burch (2009); El-Haj et al (2010); Albakour et al (2010). This is related to the technique known as *human computation*, which aims to create very large-scale resources by, for example, accumulating annotations contributed by many individuals. It is particularly appealing for creating natural language processing resources (see Chamberlain et al 2013; Poesio et al 2013, for example).

2.1.2 Automatic Creation of Language Resources

Automatic, computer-based methods have some promise of providing a more economical alternative to human-created, large-scale language resources, particularly when unsupervised methods can be applied.

Such methods are often applied to material that has been sourced from the web. For example, the WaCky corpora (Baroni et al 2009) include more than a billion words in three languages — English (ukWaC), German (deWaC) and Italian (itWaC). The corpora were built by crawling, downloading and processing web documents, and then applying basic linguistic annotation including part-of-speech tagging and lemmatisation. They are intended as general-purpose resources for the target languages. Dolan et al (2004) applied an unsupervised technique to construct a large paraphrase corpus using a simple string edit distance and a heuristic strategy to pair initial sentences from different news stories within the same collection. The PAN plagiarism corpus PAN-PC-09 used an automated process to insert 94,202 cases of artificial plagiarism in 41,223 documents (Potthast et al 2013). Grid architectures ("group of distributed computers", Foster et al 2002) have been used to process large scale data to automatically build a large scale text corpus (Li et al 2007). Chiarcos et al (2010) created a corpus of multiple parses of German sentences using a probabilistic constituent parser (BitPar) and a rule-based parser (B3 Tool) that produces semantically enriched dependency parses. They combined BitPar and B3 Tool parses in a way to provide a more reliable linguistic analysis. The purpose of this resource is to assess weaknesses in automatic parsers. Nemeskey and Simon (2012) presented a language-independent fully automated approach to build Named Entity (NE) annotated corpora from Wikipedia[10]. They used the links between the articles on Wikipedia and the in-article links (redirect and other language links) to identify entities in the sentences and tag them using a supervised "silver standard". Ptaszynski et al (2012) automatically annotated a five billion word corpus of Japanese blogs using a word and sentence level affect analysis and detailed analysis of emoticons. The annotated information includes affective features (emotive/non-emotive) and emotion classes (joy, sadness, etc.). Kozareva and Hovy (2013) implemented a minimally

[9] See also Albakour et al (2010).

[10] http://www.wikipedia.org/

supervised, automated algorithms to build terminology taxonomies and wordnets for different languages by harvesting large amounts of online domain-specific general text.

This demonstrates that it is not unusual to use automatic approaches to creating large-scale natural language resources. When the automatic techniques are language-independent, they can be adapted relatively easily to develop similar corpora for other languages. One drawback of any automated approach is the high likelihood that the resource will contain "noise" which is introduced by errors and the inevitable simplifications and approximations in the automated analysis.

2.1.3 Creation of Language Resources for Under-resourced Languages

Language such as English, German and French, have a vast number of resources and natural language processing tools. This is due to a long history of research and investment. The resources include freely available corpora and tools, such as part-of-speech taggers, semantic parsers, morphological and morphosyntactic analysers, et cetera. But many languages suffer from a lack of resources. This can impede progress on NLP with these languages. Those working with such languages often face a 'cold-start' problem.

There is some work on developing resources with other languages. Walther and Sagot (2010) proposed a methodology for building language resources from scratch. In their work they focused on Sorani Kurdish. Using a small annotated corpus and a basic seed lexicon to train a part-of-speech tagger they generated a large-scale part-of-speech tagged corpus.

Getao and Miriti (2006) constructed a Swahili (Kiswahili) corpus from the Web. They first downloaded documents from the Web and then built a Swahili language model using unigram and bigram counts. The model was used to determine if a document is in Swahili and to construct search queries to help retrieve more documents.

Guevara (2010) used the approach of Baroni et al (2009) to build a web-based corpus for Norwegian. The corpus was created by crawling, downloading and processing web documents in top level "`.no`" Internet domain. The corpus contains 700 million tokens. It would be difficult to expand beyond this due to the limits imposed by the number of public Norwegian documents on the Internet. The corpus is nearly one third the size of the published comparable Italian, German and English corpora.

Nguyen et al (2009) built a large syntactically annotated corpus for Vietnamese by constructing a treebank. To create the annotated corpus they followed the same approach that was used to create the English Penn Treebank (Marcus et al 1993): automatic parsers annotated the corpus and human annotators corrected any errors. The Vietnamese treebank consists of 10,368 annotated sentences with 210,393 words and 255,237 syllables. They also syntactically labelled 9,633 sentences with 208,406 words and 251,696 syllables.

Green et al (2012) describe the creation of the Indonesian dependency treebank. The treebank they created was a collection of manually annotated Indonesian dependency trees which consisted of 100 Indonesian sentences with 2,705 tokens and a vocabulary size of 1,015 unique tokens. They successfully trained their dependency

parser using the manually annotated data. There is hope that this will help them to provide a semi-supervised annotation and expand the corpus size.

Buhay et al (2010)'s AUTOLEX is an automatic lexicon builder intended to be used with minority languages. In their study they considered Tagalog — an Austronesian language spoken as a first language by a quarter of the population in the Philippines. They retrieved documents from the Internet written in the target language and built a lexicon using frequency-based algorithms.

2.2 Text Summarisation

Document summarisation is the process of creating an abstract or precise of a document, or collection of related documents, in order to present the key ideas in a more concise form. Manual summarisation can be time-consuming and expensive. The summaries may also be subjective. Different individuals will produce different summaries, highlighting, and obscuring, different facets. Automatic summarisation seeks to address these concerns. Work on automatic summarisation started more than 50 years ago (Luhn 1958) and is still an active area of research, as illustrated by many tools, techniques and evaluations that keep emerging (Sekine and Nobata 2003; Wang et al 2010; Nenkova and McKeown 2012).

The quality of a summary may be judged by whether it identifies appropriate aspects, whether it is readable, whether it maintains appropriate relationships, such as the chronological order of events, and whether it avoids repetition. Human precises, or abstracts, are typically written from scratch, so that they form a coherent narrative. This requires deep semantic understanding, and language generation. In contrast to such *abstractive* summaries, mechanically generated summaries are typically *extractive*: they consist of a selection of sentences taken from the document. The quality of such extractive summaries can be evaluated by using metrics to compare against gold-standard summaries, either abstractive or extractive. Without gold-standard resources for the language concerned, it can be difficult to make progress in producing high-quality automatic summarisation algorithms.

There are a range of tools and techniques that have been employed in automatic text summarisation. For extractive summaries, the tasks are to identify the key sentences. In the case of multi-document summarisation, as opposed to single-document summarisation, the problem of eliminating redundancy becomes particularly important. Fukumoto et al (2010) used clustering to eliminate redundancy where they classified the extracted sentences into groups of semantically related sentences. Hendrickx et al (2009) used semantic overlap to identify redundant sentences in their Dutch multi-document summariser system. Bossard and Rodrigues (2010) combined multi-document summarisation with a genetic algorithm (Banzhaf et al 1998). They used clustering to detect redundant sentences. The genetic algorithm was used to adapt the system to specific domains. Wang and Li (2012) developed a weighted consensus summarisation method to combine results from single-document summarisation systems. They studied different methods for multi-document summarisation, including centroid-based, graph-based, and dimension reduction to improve the summarisation quality. They used the DUC-2002 and DUC-2004 datasets. They com-

pared their method with various baselines based on average score, average rank and other measures. Barrera and Verma (2011) used a combination of syntactic and semantic approaches. They developed a single-document summariser that exploits a document's word popularity, sentence position, and semantic linkage as three main approaches for sentence extraction.

Automatic text summarisation intersects with many other natural language processing fields, including information extraction, automated question answering, natural language generation and machine translation. The overlap suggests that automatic text summarisation can be viewed as forming part of a larger picture of NLP techniques and problems.

2.3 Arabic Natural Language Processing

2.3.1 The Arabic Language

The Arabic language is the most widely spoken member of the family of Semitic languages. It is closely related to Amharic and Aramaic. Arabic is spoken by around 400 million people living in the Middle East, North Africa, and the Horn of Africa (Prochazka 2006). This suggests a large potential audience for Arabic NLP. Like other languages, literary Arabic continues to evolve. Classical Arabic (especially from the pre-Islamic to the Abbasid period, including Qur'anic Arabic) can be distinguished from Modern Standard Arabic (MSA) that is used nowadays in news and media. Arabic has several spoken dialects (varieties of Arabic). Different dialects are spoken in different countries, and in different regions of the same country. These are rarely written. This is in contrast to MSA which is mostly written and rarely spoken. In effect there is only one written language.

There has been some work on applying NLP to Arabic, and Arabic corpora (Al-Sulaiti et al 2006). There has been quite a lot of work on Arabic syntax (Sawalha and Atwell 2010b; Al-Shammari and Lin 2008; Benmamoun 2007). Yaseen and Theophilopoulos (2001) developed NAPLUS, a prototype system for processing and understanding the Arabic language. They developed utilities and tools to support Arabic NLP research in many areas such as automatic translation, text abstraction, question answering interfaces to databases and many related applications. Recently, there has been other work on syntactic annotation of Qur'anic Arabic Dukes et al (2013).[11] Kilgarriff et al (2013) developed a monolingual and bilingual word lists database, KELLY, for nine languages including Arabic. The Arabic word list contains the top 9,000 most frequent Arabic words. Buckwalter and Parkinson (2011) put together a list of the top 5,000 most frequent Arabic words, in addition the word list contains dialectical Arabic words.

As elsewhere in the world, the number of Internet users in the Arab world has been increasing (reaching 90 million by the end of 2012).[12] In the last decade, the volume of Arabic textual data has also grown on the Internet, and Arabic software for

[11] http://corpus.quran.com/
[12] http://www.internetworldstats.com/

browsing the Web has improved in quality. It can be argued that this increased access to the Internet creates a need, and an opportunity, for automatic summarisation.

2.3.2 Challenges in Arabic NLP

Like other Semitic languages, Arabic has a complex 'root-and-pattern' morphology. The root is a collection of consonants. The meaning of a word is determined by both the root and the pattern into which the root is embedded. This complicates the process of interpretation. For many languages, NLP techniques can benefit from stemming (for example, in order to reduce the dimensionality of vector-space models). But in the case of Arabic, finding the root, or stem, of a word is challenging to automate. And the root of a word often has a very abstract meaning which is not at an appropriate level for NLP. In addition, words in Arabic can be 'borrowed' from other contexts, increasing ambiguity, presenting a challenge to the mechanical interpretation of Arabic.

It seems there are some aspects about the nature of Arabic that have slowed down the progress in Arabic NLP compared to the accomplishments in English and other European languages (Diab et al 2007; Roberts et al 2006). These include the following:

- The absence of capitalisation in Arabic makes it hard to identify proper nouns, titles, acronyms, and abbreviations.
- Arabic is highly inflectional and derivational, which makes morphological analysis a very complex task.
- Diacritics (vowels) are, most of the time, omitted from the Arabic text, which makes it hard to infer the word's meaning and therefore, it requires complex morphological rules to tokenise and parse the text.

One of the challenges faced Kilgarriff et al (2013) when creating the Arabic word list was the absence of Arabic diacritics (vocalisation). They also found that the corpus for Arabic proved to have a bias towards religious terminology as a relatively high frequency of religious terms and phrases were found.

In addition to the above linguistic issues, there is also a shortage of Arabic corpora, lexicons and machine-readable dictionaries. These tools are essential to advance research in different areas. Arabic NLP has focused on the morphological, lexical and syntactic properties of Arabic. Semantic processing of the Arabic language — even broadly construed — is still in its early stages (Haddad and Yaseen 2005; Diab et al 2007; Al-Shammari and Lin 2008; Hmeidi et al 2010; Benajiba et al 2010). Morphological analysis of words in a text is the first stage of most natural language applications. It has been regarded as being particularly crucial in processing highly inflectional and derivational languages like Arabic (Diehl et al 2012). Although Arabic morphology is considered complex compared to many other languages, attempts to tackle this problem have been quite successful (Sawalha and Atwell 2010a; Beesley 1998; Abuleil et al 2002; Darwish et al 2005; Smrž 2007; Habash and Roth 2011). There are also a number of tools for tokenisation and stemming of Arabic (Larkey et al 2002; Al-Ameed et al 2006; Attia 2007; Al-Shammari and Lin 2008; Hmeidi et al 2010; Althobaiti et al 2014).

Some commercial Arabic resources and tools are available online, such as Sakhr[13] and RDI[14]. There has been an increased demand for resources and tools to assist and advance the research on Arabic. There are also some annotated Arabic text treebanks (such as the Arabic Penn Treebank and Prague Arabic Dependency Treebank). There has been some progress in addressing limited Arabic resources for specific areas, such as plagiarism detection, for example (Bensalem et al 2013).[15] But gold-standard collections for Arabic text summarisation have not been available until we started work in this area. Arabic has not featured in popular competitions such as TAC and DUC.

While the absence of Arabic resources for summarisation initially presented us with a problem, in the end it provided an opportunity to explore different methods for developing resources for a widely used but relatively under resourced language that could provide lessons for other languages that currently lack resources.

3 Three Approaches to Create Language Resources

Different language resources are required for different NLP applications. For evaluation purposes we refer to them as 'test collections' which typically include appropriate gold standards. Measuring the performance of different techniques against gold standards allows the different results to be compared. For automatic summarisation, test collections are required that contain documents with gold-standard summaries. The summaries may be human-generated, either by human experts or some form of crowdsourcing. In the case of the cross-linguistic approach that we tried, such summaries may also be translated into other languages using human translators and machine translation tools.

For languages such as English resources with gold-standard summaries are readily available. Furthermore, numerous results have been published. This allows researchers to compare their work with that of others when judging their summariser's quality and performance. This is not the case for all languages. In particular, we found there were few appropriate and readily-available resources for Arabic summarisation to evaluate our results. Nor were there published results with which our work could be compared directly. This meant we had to find ways of producing appropriate resources quickly and cheaply, and find novel ways of comparing our results to state-of-the-art summarisation for other languages.

To create resources for single and multi-document text summarisation we need to identify an appropriate collection of documents, whose use is not restricted for such work, and find fluent users of the language who are able to help build the resource, such as being able to produce and verify gold-standard summaries. Ideally the resource should be made available to a wider audience to allow others to publish their results as measured against the gold-standard. One obstacle is then finding documents whose copyright terms allow them to be distributed to others. Another is

[13] http://www.sakhr.com/

[14] http://www.rdi-eg.com/technologies/arabic_nlp.htm

[15] Other languages also suffer from a lack of resources for plagiarism detection, including Basque for example (no et al 2010).

finding appropriately skilled participants and involve them in a way that is not too time-consuming or costly. This problem is more difficult if there is little funding.

For both multi-document and single-document summarisation, datasets can be obtained from websites that offer relatively liberal copyright terms.[16] For multi-document summarisation, any dataset selected should ideally contain groups of related articles. An example would be news articles about a topic from multiple sources.[17]

A critical issue is finding skilled users of the language. The participants' roles can include creating manual summaries and evaluating them. It is appropriate to have a number of *model summaries* for each document or set of related articles, each created by a different participant. This helps reduce the impact of individual bias in the summaries. As we were exploring cross-linguistic approaches to evaluation, we also required participants who could translate summaries, and evaluate the quality of translations.

We created resources for Arabic single and multi-document summarisation with gold-standard summaries. As will be seen, the tasks involved included summarising Arabic text documents, translating articles from English into Arabic, summarising the translated articles, and evaluating both the translations and the summarisation quality. These tasks were performed manually and with the aid of computer tools.

This work provides guidance for others who need to create resources for under-resourced languages. Furthermore, the resources themselves have been made available to the research community with the objective of helping to advance the state of Arabic NLP, especially single and multi-document summarisation.

3.1 Crowdsourcing Language Resources

We first describe how we used crowdsourcing to construct a single-document summarisation corpus for Arabic. Crowdsourcing allowed us to do this rapidly and relatively cheaply.

This can be seen as an initial pilot. It was our starting point before moving on to the creation of corpora for multi-document summarisation.

We did not consider creating a multi-document summaries corpus using crowdsourcing. A key challenge was the problem of determining an objective way of handling potential "spam" answers for this more demanding task.[18] In this context, Lloret et al (2013) found that crowdsourcing might not be the right service to generating informative summaries from a set of multiple related documents.

Someone could think of creating summaries from Wikipedia using just the first sentence as the summary. Even though the first sentence has been shown to be a

[16] An example of a large website with liberal copyright terms is Wikipedia (http://www.wikipedia.org/). It is important to note that corpora drawn from such sites still need to ensure that the terms of the copyright are being followed.

[17] An example of such a source is the Wikinews website (http://www.wikinews.org/). To illustrate a potential problem, we were granted the use of news articles from a large UK-based website, but our rights did not permit further distribution of those articles. These articles had to be removed from the public versions of our datasets.

[18] There are some Bayesian techniques that might be adapted to help address this issue, but that lay outside the scope of our work (for example, Carpenter 2008, and related work).

significant feature in many summarisers (Fattah and Ren 2008; Yeh et al 2008), it is interesting to note that summaries consisting of a single sentence only (e.g. Baseline–1 in El-Haj et al 2010) do not score particularly well, which suggests that the first sentence is important but not sufficient for a good summary. Instead, we used Amazon's Mechanical Turk (AMT) to recruit a sufficient number of fluent users of Arabic to create our gold-standard summaries of articles obtained from Wikipedia and Arabic-language news sources. This work lead to the *Essex Arabic Summaries Corpus (EASC)*. The corpus includes manually created summaries. These are intended to be used as a gold standard for evaluating Arabic summarisers. Some statistics for EASC are given in Table 1. Below, we give the details of how this corpus was created. Figure 1 and 2 show a sample document and its summary created by AMT.

العود هي آلة موسيقية شرقية وترية تاريخها موغل بالقدم يرجعه البعض إلى نوح. و تعني كلمة العود في اللغة العربية الخشب. فالعود من الآلات الوترية العربية له خمسة أوتار ثنائية أو و يمكن ربط وتر سادس إلى العود و يغطي مجاله الصوتي حوالي الأوكتافين و نصف الأوكتاف. الأدبيات الشرقية الموسيقية كلها تظهر وتؤكد على استخدام نوع أو أكثر من الأعواد. يعتبر العود آلة رئيسية في التخت الموسيقي الشرقي, في التاريخ الحديث هناك أكثر من دولة عربية تدعي تفوقها في صناعة الأعواد ولكن بغداد لها السبق في ذلك ففي بغداد هناك أكثر من حرفي يقوم بصنعها, وبذلك غدا العود العراقي ذا سمعة عالمية جعلت كبار الملحنين والمطربين في العالم العربي يفضلونه على غيره, ومن أحد صناع العود المحدثين في العراق سمير رشيد العواد الذي يتسابق الفنانون العرب للحصول على أحد الأعواد من صنعه, مثل المطرب الكويتي نبيل شعيل ، والمطرب اللبناني الكبير وديع الصافي والمطرب العراقي المشهور كاظم الساهر والمطرب مهند محسن والمطرب ياس خضر والموسيقار منير بشير.

Fig. 1 EASC: AMT Document Sample *(EASC Corpus: Topic 3 – Document: Art and Music (3).txt)*

العود هي آلة موسيقية شرقية وترية تاريخها موغل بالقدم يرجعه البعض الى نوح. الادبيات الشرقية الموسيقية كلها تظهر وتؤكد على استخدام نوع او اكثر من الاعواد.

Fig. 2 EASC: AMT Summary Sample *(EASC Corpus: Topic 3 – Summary: D0103.M.250.A.1.A)*

Table 1 Statistics for the Essex Arabic Summaries Corpus (EASC)

Documents:	153
Sentences:	2,360
Words:	41,493
Distinct Words:	18,264
Gold-standard summaries:	765
Summaries per document:	5

3.1.1 The Document Collection

The document collection used in the creation of the single-document summaries corpus was extracted from the Arabic language version of Wikipedia and two Arabic newspapers; Alrai[19] (published in Jordan) and Alwatan[20] (published in Saudi Arabia). These sources were chosen for the following reasons:

1. They contain real text as written and used by native speakers of Arabic.
2. They are written by different authors from different backgrounds.
3. They cover a range of topics from different subject areas, each with a credible amount of data.
4. We had the copyright holders' permission to distribute the results.[21]

The Wikipedia documents were selected by a group of students. They were asked to search for Wikipedia articles within ten given subject areas (art and music; the environment; politics; sports; health; finance and insurance; science and technology; tourism; religion; and education). To obtain a more uniform distribution of articles across topics, the collection was then supplemented with newspaper articles. These were retrieved from the newspapers' websites using the same queries that the students used to select the Wikipedia articles. Overall, a total of 153 documents were used, containing a total of 18,264 words. On average, each document consisted of 380 words, with a minimum of 116 words and a maximum of 971 words.

3.1.2 Creating Manual Summaries

The extractive summaries for the documents were generated by way of *Human Intelligence Tasks* (HITs) published on AMT. Any individual participating in one of these tasks was required to produce a summary of a given document by selecting those sentences they thought were the most important/significant. The sentences in the document were displayed to the participant as an enumerated list. The participant then used the numbers to identify those sentences they believed should be in

[19] http://www.alrai.com/

[20] http://www.alwatan.com.sa/

[21] Originally it was planned to include news articles from the BBC website (http://www.bbc.co.uk/news/). We obtained permission to use these articles for our work on summarisation. Unfortunately the terms of use were such that it would have been impractical to distribute the corpus to others. For this reason, these articles and associated summaries were excluded from the final corpus.

the summary. They were required to select no more than half of the sentences in the article.[22]

Using this method, five summaries were created for each article in the collection. Each of the summaries for a given article was generated by five different individuals. In the case of annotation tasks, it has been demonstrated that aggregating multiple independent annotation from different workers can produce good quality results (Albakour et al 2010; Kazai et al 2011). We assumed that this held true for the sentence selection task.

Participants were required to complete a quality control task. This was aimed at ensuring participants were properly engaged with the content of the articles, rather than selecting sentences "at random". To this end, participants were asked to provide up to three keywords for the article they were summarising. To avoid the risk of introducing subjective bias, the results of this task were not used to filter the final results. For the same reason, even apparently idiosyncratic sentence selections are included.

3.1.3 Creating Gold-standard Summaries for EASC

The gold-standard summaries for the corpus were generated by combining the participants' sentence selections. Essentially, the selection of a sentence from a document was treated as a "vote" in favour of including that sentence in a gold-standard summary. The number of votes for a sentence had to pass a threshold for it to be included in a gold-standard summary. This amounts to an aggregation method for combining the results (Kittur et al 2011).

To obtain a better understanding of the impact of the aggregation method, we selected three different thresholds, giving rise to three different gold-standard summaries for each document. At the lowest threshold ("*Level 1*", or "*All*"), a sentence was included in a gold-standard summary of a document if any of the participants had selected it. For the intermediate threshold ("*Level 2*"), a sentence had to be selected by at least two participants for it to be included in the gold-standard summary. And for the highest threshold ("*Level 3*") a sentence had to have been selected by at least three of the five of the participants.[23]

The *Level 1* summaries are likely to contain a lot of noise in the form of idiosyncratic outlier sentences. For this reason only the *Level 2* and *Level 3* summaries should be regarded as providing genuine gold-standard summaries, with reduced subjective bias. The summaries from *Level 1* merely give us a point of comparison.

[22] Appendix A shows the guidelines given to the workers for completing the task in addition to a HIT example. Payments made to the users were dependent on the document size, ranging from £0.04 to £0.33 per task with an estimate overall cost of £200.

[23] To illustrate the thresholds, assume we provided three participants with a document made of six sentences. Following the EASC summarisation guidelines in Appendix A, the users must provide a selection of up to three sentences as a summary. For example, consider the following selections, where A, B and C are three random participants and the numbers identify the sentences they selected: *A*(1,3,5), *B*(1,2,3), *C*(1,4,5). The three gold-standard summaries will be as follows: the *Level 3* summary will consist of sentence 1 only, *Level 2* summary will contain sentences 1, 3 and 5. And finally the *All* summary will contain sentences 1,2,3,4 and 5.

3.1.4 Summary

In addition to producing a useful resource for Arabic NLP, this work also demonstrated the application of inexpensive crowdsourcing for natural language resource creation. The documents and 765 human-generated summaries of EASC are freely available to the community.[24] The size of the corpus was determined in part by the available financial resources. The method can be scaled quite easily if more resources are available. The method used to produce EASC could be adopted, and adapted, to produce a significant body of gold-standard single document summaries and other NLP resources for a range of languages. The actual process is language independent, provided that a sufficiently large number of users of the language are available.

3.2 Creation of Resources using Machine Translation

The second approach we illustrate here is to translate an existing gold-standard resource into the target language. Similar to Aker and Gaizauskas (2010) where they used machine translation to translate summaries in English into German, here we describe using machine translation to create an Arabic corpus of multi-document summaries from an English language dataset without any post-editing on the translated summaries. The English dataset chosen was from DUC-2002.[25]

There are numerous published results that use the DUC-2002 dataset as a benchmark for multi-document summarisation. This provides us with performance figures that can be used when evaluating the results of Arabic multi-document summarisation on the translated corpus. This was an important factor in our decision to construct a translated version of the DUC-2002 dataset.

The DUC-2002 dataset contains 567 articles and 1,111 gold-standard (model) summaries. The documents are grouped into "reference sets" — a reference set being a small collection of documents about a particular topic. These reference sets provide appropriate source material for multi-document summarisation. In this case, the documents in a reference set all fell into one of the following four categories, with originally fifteen reference sets for each category.[26]

1. Documents discussing a single natural disaster created within a seven day window.
2. Documents about a single event in any domain and created within at most a seven day window.
3. Documents about multiple distinct events of a single type with no limit on the time window.

[24] EASC can be obtained from `http://sourceforge.net/projects/easc-corpus/`. To simplify the use of EASC when evaluating summarisers, the file names and extensions are formatted to be compatible with evaluation systems such as ROUGE (Lin 2004) and AutoSummENG (Giannakopoulos et al 2008). It is also available in two character encodings, UTF-8 and ISO-8859-6 (Arabic).

[25] The DUC-2002 dataset was as provided by the National Institute of Standards and Technology (NIST — `http://www.nist.gov/index.html`) through the Document Understanding Conference (DUC).

[26] See `http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html`. Note that there are some discrepancies in the published statistics for the corpus: NIST withdrew some of the documents, leaving just 59 reference sets rather than the original 60.

4. Documents presenting biographical information mainly about a single individual.

The reference sets were produced using data from the TREC question-answering track in TREC-9[27]. Table 2 shows the corpus statistics of the DUC-2002 Arabic translation. Figure 3 shows a sample summary created by the use of machine translation technology.

إعصار جيلبرت جلب أكثر من 24 ساعة من الامطار الغزيرة على مونتيري وماتاموروس ، والمكسيك ، مما تسبب في فيضانات شديدة وحالة وفاة في المنطقة الصناعية في مدينة مونتيري ، وضباط الشرطة على الأقل 10 في مهمة الانقاذ التي جرفتها نهر سانتا كاتارينا.

Fig. 3 Machine Translation Summary Sample *(DUC 2002 Data: a summary of the Arabic translation of English reference set d079a)*

Table 2 Statistics for the DUC-2002 corpus as translated into Arabic

Documents:	567	
Sentences:	17,340	
Words:	199,423	
Distinct words:	19,307	
Sentences per document:	≥ 10	
Reference sets:	59	
Documents per reference set:	5–15	(typically ten)
Multi-document extractive summaries:	118	(two per reference set)

The model summaries included both single and multi-document abstractive and extractive summaries. In this case we are only interested in the multi-document extractive summaries. Each reference set had two multi-document extractive summaries, one of at most 400 words, and another of at most 200 words. No partial sentences were used. The sentences in the shorter summary all appeared in the larger summary.

3.2.1 Translation of DUC-2002 Dataset

The DUC-2002 dataset was translated automatically into Arabic sentence-by-sentence using the Java version of Google Translate API.[28] A total of 17,340 sentences had to be translated. Google Translate limits the size of text and the number of requests that can be made in a given period. This was accommodated by translating only one sentence every half-second.

This approach to translating a gold-standard resource can be applied to other languages, provided that an appropriate translation system is available for the specific

[27] http://trec.nist.gov/pubs/trec9/t9_proceedings.html
[28] http://code.google.com/p/google-api-translate-java/

language pair. The method should be helpful when there is a shortage of human participants who know the target language, or when there is little or no funding available. It also has the advantage of allowing those working in the target language to compare their results with those obtained by state-of-the-art systems in the source language, on the very same dataset.

3.3 Creation of Language Resources using Human Experts

The third paradigm we illustrate involves the use of selected human participants to create a resource for Arabic multi-document summarisation, with no automation or anonymous "crowdsourcing". The tasks included translation, summarisation and validation.

This corpus was created through our involvement in the organisation of the TAC-2011 MultiLing Summarisation Pilot.[29] The test collection consisted of 100 documents drawn from WikiNews texts.[30] The source documents were in the form of plain text UTF-8 encoded files, with no meta-data or tags. The 100 documents were assigned to ten reference sets. Each reference set contained ten related articles discussing the same topic. The original language of the dataset was English. This was translated and summarised in six languages — Arabic, Hindi, French, Czech, Greek and Hebrew — to give a multi-lingual test collection.

The preparation of the Arabic corpus for the TAC-2011 MultiLing Summarisation Pilot is our focus of interest here. Table 3 shows the corpus statistics of the TAC-2011 Arabic dataset.

Table 3 Statistics for the TAC-2011 Arabic Corpus

Documents:	100
Sentences:	1,573
Words:	30,908
Distinct words:	9,632
Reference sets:	10
Documents per reference set:	10
Gold-standard summaries:	30
Summaries per reference set:	3

A total of twelve people participated in creating the Arabic version of the corpus. They were selected based on their proficiency in Arabic and English. Each one of them had to have Arabic as a first language. The participants were either studying, or had completed a university degree in an Arabic-speaking country. The participants were between 22 and 64 years of age. The participants were paid using Amazon vouchers. The value of the vouchers awarded to each participant depended on the task performed. The total value of Amazon vouchers paid to the participants was £250.

[29] http://www.nist.gov/tac/2011/Summarization/index.html
[30] http://www.wikinews.org/

The participants translated the English dataset into Arabic. For each translated article another translator validated the translation and fixed any errors. Three extractive summaries were manually created for each of the translated reference sets.[31] Each of the summaries for a reference set were created by different participants. Participants in the summarisation process were required to evaluate the quality of other summaries. They did so by assigning a score between one (unreadable summary) and five (fluent and readable summary). No self-evaluation was permitted. Appendix B shows the detailed guidelines given to the participants.

The manual evaluation was based on the *Overall Responsiveness* of a text (Appendix B). Figure 4 shows a sample summary.

أعلن رسمياً عن الانتهاء الرسمي للألعاب الأولمبية في بكين. وكانت الشعلة الأولمبية واجهت العديد من العثرات و الاحتجاجات في طريقها من سان فرنسيسكو إلى بكين مما منع من حملها باليد في الشوراع بارغم من الاجراءات الأمنية المشددة التي اقامتها السلطات في المدينة, الامر الذي انتقد من العديد من الرياضيين وأساء إلى الألعاب الأولمبية. في حين أصرت الصين أن الشعلة لا يمكن أن تأتي إلى بكين عبر تايوان مما فاقم الوضع و عبر عن تداخل السياسة بالرياضة.وقد تم الابلاغ عن انطفاء الشعلة الأولمبية لثلاث مرات في فرنسا بسبب المتظاهرين المؤيدين لاستقلال التيبت. وقد حضر حفل الافتتاح ما يقارب 90000 متفرج على الاستاد الوطني في بكين. وكانت اول مقولة في الافتتاح "كم نحن سعداء بلقاء الأصدقاء من جميع أنحاء العالم. وكان الأولمبياد قد استمر ل 16 يوما و شاركت فيه 214 دولة من مختلف أنحاء العالم تنافست على 38 لعبة و 302 ميدالية. حيث أن الصين تصدرت لائحة الميداليات ب 51 ميدالية ذهبية و جاءت الولايات المتحدة في المرتبة الثانية ب 36 و روسيا في المركز الثالث ب 23. ومما اشتهرت به الألعاب في بكين بعد امكانية قيادة السيارات الخاصة للممتابعين للحدث بسبب الازدحام الشديد في هذه المدينة. وقد تم دعوة جميع الرياضيين و الاشخاص في جميع انحاء العالم إلى الاستعداد للأولمبياد القادمة بعد أربع سنوات عام 2012 في لندن. وبعد اختتام الألعاب من المتوقع أن تؤخد الشعلة ما يقارب مسافة 31 كيلو متر في جميع أننحاء لندن بداية من ملعب ويمبلي اشارة إلى بدأ التحضيرات لاستقبال العاي عام 2012. ومن المتوقع أن يصطف العديد من الأشخاص في الشوارع لمتابعة هذا الحدث.

Fig. 4 TAC 2011 MultiLing: Human Expert Summary Sample *(MultiLing 2011 Data: Arabic summary of reference set M009 by human summariser B)*

The average time for reading the English news articles by the Arabic native speaker participants was four minutes. The average time it took them to translate those articles into Arabic was 25 minutes, and to validate each of the translated Arabic articles the participants took six minutes on average. For the summarisation task the average time for reading the set of related articles (ten articles per each set) was 17 minutes. The average time for creating a multidocument summary of a reference set was 24 minutes.

Until TAC-2011, participation to DUC and later TAC workshops was essentially limited to English summarisation. The creation of the TAC-2011 MultiLing corpora can be seen as an example of addressing the need for more gold-standard resources

[31] Although the gold-standard summaries were extractive in nature, the TAC-2011 MultiLing Summarisation Pilot allowed participating systems to use other approaches.

for under-resourced languages, more specifically the creation of high-quality gold-standard multi-lingual summaries for multi-document summarisation. More generally, the TAC-2011 corpus could help those working on automatic summarisation, question answering, and automatic translation. The corpus is freely available.[32]

One issue encountered in developing this resource was the difficulty in finding appropriately skilled users of the language concerned. This is a particularly acute problem when the research is based in countries where the majority of people use a different language. In our case it was difficult to find participants for whom Arabic was a first language. Most of our participants were located in other countries. We communicated with the participants online, which complicated the management of the process.

4 Evaluation

In this section we present human evaluations of some of the language resources we created, including the summaries generated using human experts and crowdsourcing. The evaluations were performed by Arabic native speakers who assessed the summaries on a five-point Likert scale. We applied a comparable evaluation methodology to both cases. Pairwise t-tests (with $p < 0.05$) have been applied were appropriate to test for significantly different results.

The summaries generated by machine translation (Section 3.2) were not evaluated. The main purpose of using this technique was to automate the evaluation of a non-English summariser, rather than to generate good quality summaries directly. In addition this would require us to evaluate the quality of automatic translation, which we considered to fall outside the scope of our work. In the case of the human generated translations, evaluation was already incorporated into the validation procedures (Appendices A, B and C).

4.1 Evaluating Crowdsourcing Language Resources

To evaluate the single-document summarisation corpus (EASC, Section 3.1) we asked three individuals (X, Y and Z in Table 4) to evaluate the quality of summaries obtained by a range of methods on a random sample of ten documents. Each participant evaluated a number of 110 summaries. The evaluators were paid £10 per hour. The evaluation task was estimated to be three hours in total.

Three crowdsourcing methods of obtaining a document summary were assessed, namely:

(1) *Crowdsourced Summaries* — randomly selecting individual summaries as submitted by MTurkers (Section 3.1.2);
(2) *Level 2* — aggregated summaries over all five individual summaries obtained for a document by applying the "Level 2" aggregation approach as explained in Section 3.1.3;

[32] The corpus can be downloaded directly after completing the relevant forms, as required by the MultiLing organisers (`http://multiling.iit.demokritos.gr/file/all/`).

(3) *Level 3* — similar to "Level 2" but with stricter aggregation (Section 3.1.3).

In addition, four automatically generated summaries were also evaluated:

(4) *Baseline–1* — where the summary consists of the first sentence of each document;
(5) *Baseline–2* — the first two sentences of each document;
(6) *Baseline–3* — the first three sentences of each document; and
(7) a *centroid* summary (see Radev et al 2000, 2004; Sarkar 2009, for example).

The first-sentence baseline had been occasionally used as a sensible baseline. It is similar to the first-*n*-words baseline summariser. Both of these baselines have been widely used in the DUC and TAC summarisation tasks (Katragadda et al 2009; Nenkova 2005). We included centroid summaries because they scored highly when evaluated in previous experiments (El-Haj et al 2011a,c). These summaries consist of those sentences that are closest to the vector-space centroid of the document collection (using cosine similarity), up to a maximum of 250 words, or no more than 50% of the document's sentences, whichever is the lowest. The centroid is calculated by treating the document collection as a bag of words.

The evaluation was anonymised. The evaluators were not aware whether they were evaluating human or system summaries.

Table 4 shows the results of the evaluation of the seven approaches. Baseline–1 performs significantly worse than any of the crowdsourced approaches (i.e. any approach that involved user assessments) as well as the centroid-based approach.

The centroid-based baseline performed well. As this is a fully-automated approach it is intuitively appealing as it does not involve additional costs to run experiments.

Table 4 Crowdsourcing Evaluation Scores

Summariser	**X**	**Y**	**Z**	**Mean**
Centroid Summariser	3.00	3.00	1.70	2.57
Level 2	2.60	2.70	2.10	2.47
Crowdsourced Summaries	2.32	2.48	2.50	2.43
Level 3	2.40	2.60	1.90	2.30
Baseline–3	2.10	2.90	1.60	2.20
Baseline–2	1.70	2.50	1.00	1.73
Baseline–1	1.20	1.30	1.10	1.20

One reason that *Baseline–1* does not score particularly well could be that while the first sentence may be important, it is not sufficient for a good summary (Katragadda et al 2009). Adding more sentences improves the overall assessed quality (in some cases significantly) but neither *Baseline–1* nor *Baseline–2* outperform any of the other methods.

In the case of the Level 3 summaries, this method results in relatively short summaries. For this reason this approach might suffer from a similar problem as *Baseline-1* in that the aggregated result tends to contain text that is appropriate but not sufficient for a good summary.

All other methods perform on par with each other with only marginal differences: we found no significant difference between the top performing methods.

Table 5 MultiLing Human Experts Evaluation Scores

System	Mean Evaluation
B	4.04
ID9	3.73
ID10	3.20
C	2.23
A	1.92

This evaluation only considers a sample of documents, and involves a very small number of experts. The variation of assessments given between experts is quite large. Using more assessors and documents may help identify statistically significant differences.

4.2 Evaluating Language Resources by Human Experts

We also evaluated the language resources created by the human experts for the TAC-2011 MultiLing Summarisation Pilot (Section 3.3). The same human experts mentioned in Section 3.3 evaluated 30 Arabic multi-document summaries. The manual evaluation was based on the Overall Responsiveness of a text (Appendix B). Twelve participants evaluated summaries generated by three different human experts (A, B and C) in addition to the global topline (**ID10**) and baseline (**ID9**) systems.

The global baseline system — **ID9** — is effectively a version of the centroid summariser (Radev et al 2000, 2004; Sarkar 2009) as outlined in Section 4.1. The summariser sorts sentences based on their cosine similarity to the centroid of a cluster, then starts adding sentences to the summary, until it either reaches 250 words, or it reaches the end of the document. In the latter case, it continues with the next document in the sorted list.

The global topline system — **ID10** — used information from the model summaries (i.e. cheats). First, it split all source documents into sentences. Then it uses a genetic algorithm to generate summaries that have a vector with maximal cosine similarity to the centroid vector of the model summary texts (see the "MeMoG" method, Giannakopoulos and Karkaletsis 2011, for more details).

Table 5 shows the evaluation scores of the baseline, topline systems in addition to the summaries created by the human experts.

As before, the centroid-based method (**ID9**) performs very well.

The human evaluators ranked summaries by human expert B to be of highest quality. Summaries by baseline (ID9) were ranked high by the evaluators. We assume this is because they contained continuous text, resulting in more readable summaries.

The guidelines for those producing summaries did not indicate whether abstractive or extractive summaries were required. Notes made by participants showed that they tended to use the extractive method as it is less time consuming. The summaries concatenate sentences taken from different contexts, and as a result they often lack coherence.

Those producing the summaries also expressed concern that the 250-word limit was too low when summarising ten related documents. Some of the documents could

not be represented in the summary. The guidelines were that a Likert score of five means the summary covers all the important aspects of the ten documents. As a consequence, some evaluators never awarded a score greater than four on the Likert scale. This potentially explains why baseline ID9 and topline ID10 received higher scores than A and C human summaries: the algorithms tended to pick sentences from a broader sample of the documents in each collection than the human summarisers.

5 Conclusions

Resource creation plays an important role in advancing Natural Language Processing tools and applications. Lack of resources can be a serious problem particularly for evaluation purposes, when it is important to have some external metric against which to assess different methods and technologies.

This paper illustrates three different ways of creating gold-standard resources for under-resourced languages rapidly and relatively cheaply. Our examples were concerned with corpora for Arabic-language text summarisation. The methods we illustrate were as follows:

1. *Using crowdsourcing*. Such a resource is relatively cheap and small but can be built rapidly;
2. *Automatic machine translation* of an existing gold-standard collections. This is relatively easy, but the resource is likely to be more noisy.
3. *Manual creation of a resource* using proficient users of the language(s) concerned. This is more expensive, but can result in high quality.

The use of automated methods for creating language resources is appropriate when the tools required are more readily available than appropriately qualified participants, or when no funding is available to identify and reimburse participants. The results of the evaluation suggest that it is appropriate to include centroid-based summaries as a sensible baseline; in our experiments their quality appears to be on a par with manual summarisation.

Generally our evaluations have low average scores for summaries regardless of the method. This could indicate two things: (i) the difficulty of the task of generating a good summary in general, and (ii) the limitations of an "extractive" summary as opposed to an abstractive summary. This warrants further investigations.

All of the resources described in this paper are freely available to the research community. The corpora that have been created are of use for Arab-language researchers working on automatic summarisation, question answering, and automatic translation.

References

Abouenour L, Bouzoubaa K, Rosso P (2013) On the evaluation and improvement of arabic wordnet coverage and usability. Language Resources and Evaluation 47(3):891–917

Abuleil S, Alsamara K, Evens M (2002) Acquisition system for Arabic noun morphology. In: Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages, Association for Computational Linguistics, Stroudsburg, PA, USA, SEMITIC'02, pp 1–8

Aker A, Gaizauskas RJ (2010) Model summaries for location-related images. In: The 7th International Language Resources and Evaluation Conference (LREC 2010)., LREC 2010, Valletta, Malta, pp 3119–3124

Aker A, El-Haj M, Kruschwitz U, Albakour D (2012) Assessing crowdsourcing quality through objective tasks. In: 8th Language Resources and Evaluation Conference, LREC 2012, Istanbul, Turkey

Al-Ameed H, Al-Ketbi S, Al-Kaabi A, Al-Shebli K, Al-Shamsi N, Al-Nuaimi N, Al-Muhairi S (2006) Arabic light stemmer: A new enhanced approach. In: The 2nd International Conference on Innovations in Information Technology, IIT'05, Dubai, United Arab Emirates

Al-Shammari E, Lin J (2008) Towards an error-free Arabic stemming. In: Lazarinis F, Efthimiadis E, Vilares J, Tait J (eds) Proceeding of the 2nd ACM workshop on Improving Non English Web Searching, iNEWS 2008, Napa Valley, California, USA, October 30, 2008, ACM, pp 9–16

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

Albakour M, Kruschwitz U, Lucas S (2010) Sentence-level attachment prediction. In: Cunningham H, Hanbury A, Rüger S (eds) Advances in Multidisciplinary Retrieval, Springer Berlin / Heidelberg, Lecture Notes in Computer Science, vol 6107, pp 6–19

Alghamdi M, Chafic M, Mohamed M (2009) Arabic language resources and tools for speech and natural language: KACST and Balamand. In: The 2nd International Conference on Arabic Language Resources and Tools, Cairo, Egypt

Alonso O, Mizzaro S (2009) Can we get rid of TREC assessors? using Mechanical Turk for relevance assessment. In: SIGIR '09: Workshop on The Future of IR Evaluation

Althobaiti M, Kruschwitz U, Poesio M (2014) AraNLP: a Java-Based Library for the Processing of Arabic Text. In: Proceedings of the 9th Language Resources and Evaluation Conference (LREC), Reykjavik

Attia M (2007) Arabic tokenization system. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Association for Computational Linguistics, Stroudsburg, PA, USA, Semitic '07, pp 65–72

Banzhaf W, Francone F, Keller R, Nordin P (1998) Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

Baroni M, Bernardini S, Ferraresi A, Zanchetta E (2009) The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation (LREV) 43(3):209–226, DOI 10.1007/s10579--009--9081--4

Barrera A, Verma R (2011) Automated extractive single-document summarization: Beating the baselines with a new approach. In: Proceedings of the 2011 ACM Symposium on Applied Computing, ACM, TaiChung, Taiwan, SAC'11, pp 268–269

Beesley K (1998) Arabic morphology using only finite-state operations. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages, Association for Computational Linguistics, Stroudsburg, PA, USA, Semitic '98, pp 50–57

Benajiba Y, Diab M, Rosso P (2009) Arabic named entity recognition: A feature-driven study. Audio, Speech, and Language Processing, IEEE Transactions on 17(5):926–934

Benajiba Y, Zitouni I, Diab M, Rosso P (2010) Arabic named entity recognition: using features extracted from noisy data. In: Proceedings of the ACL 2010 conference short papers, Association for Computational Linguistics, pp 281–285

Benmamoun E (2007) The syntax of Arabic tense. Cahiers de Linguistique de L'INALCO 5:9–25

Bensalem I, Rosso P, Chikhi S (2013) A new corpus for the evaluation of arabic intrinsic plagiarism detection. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds) Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Lecture Notes in Computer Science, vol 8138, Springer Berlin Heidelberg, pp 53–58

Bossard A, Rodrigues C (2010) Combining a multi-document update summarization system "CBSEAS" with a genetic algorithm. In: International Workshop on Combinations of Intelligend Methods and Applications, Hyper Articles en Ligne, Arras, France, CIMA 2010

Boyer A, Brun A (2007) Natural language processing for usage based indexing of web resources. In: Amati G, Carpineto C, Romano G (eds) Advances in Information Retrieval, Lecture Notes in Computer Science, vol 4425, Springer Berlin Heidelberg, pp 517–524, DOI 10.1007/978-3-540-71496-5_46, URL http://dx.doi.org/10.1007/978-3-540-71496-5_46

Buckwalter T, Parkinson D (2011) A Frequency Dictionary of Arabic: Core Vocabulary for Learners. Routledge Frequency Dictionaries, Routledge, URL http://books.google.co.uk/books?id=Kj_NRwAACAAJ

Buhay E, Evardone M, Nocon H, Dimalen D, Roxas R (2010) Autolex: An automatic lexicon builder for minority languages using an open corpus. In: PACLIC'10, pp 603–611

Callison-Burch C (2009) Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 — Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '09, pp 286–295

Calzolari N, Soria C, Gratta RD, S Goggi VQ, Russo I, Choukri K, Mariani J, Piperidis S (2010) The LREC 2010 resource map. In: The 7th International Language Resources and Evaluation Conference (LREC 2010), LREC 2010, Valletta, Malta, pp 949–956

Carpenter B (2008) Multilevel bayesian models of categorical data annotation. Available at http://lingpipe-blog.com/lingpipe-white-papers

de Chalendar G, Nouvel D (2009) Modular resource development and diagnostic evaluation framework for fast nlp system improvement. In: Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, SETQA-NLP '09, pp 65–73

Chamberlain J, Fort K, Kruschwitz U, Lafourcade M, Poesio M (2013) Using games to create language resources: Successes and limitations of the approach. In: The People's Web Meets NLP, Theory and Applications of Natural Language Processing, Springer, pp 3–44

Chiarcos C, Eckart K, Ritz J (2010) Creating and exploiting a resource of parallel parses. In: Proceedings of the Fourth Linguistic Annotation Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, LAW IV '10, pp 166–171

Darwish K, Hassan H, Emam O (2005) Examining the effect of improved context sensitive morphology on Arabic information retrieval. In: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Association for Computational Linguistics, Stroudsburg, PA, USA, Semitic '05, pp 25–30

Diab M, Hacioglu K, Jurafsky D (2007) Automatic processing of modern standard Arabic text. In: Soudi A, van den Bosch A, Neumann G (eds) Arabic Computational Morphology: Knowledge-based and Empirical Methods, Text, Speech and Language Technology, Springer Netherlands, pp 159–179

Diehl F, Gales M, Tomalin M, Woodland P (2012) Morphological decomposition in Arabic ASR systems. Computer Speech and Language 26(4):229–243

Dolan B, Quirk C, Brockett C (2004) Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th International Conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '04

Dukes K, Atwell E, Habash N (2013) Supervised collaboration for syntactic annotation of Quranic Arabic. Language Resources and Evaluation Journal (LREV) 47(1):33–62, special Issue on Collaboratively Constructed Language Resources

El-Haj M, Kruschwitz U, Fox C (2010) Using Mechanical Turk to create a corpus of Arabic summaries. In: Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages workshop held in conjunction with the 7th International Language Resources and Evaluation Conference (LREC 2010), LREC 2010, Valletta, Malta, pp 36–39

El-Haj M, Kruschwitz U, Fox C (2011a) Exploring clustering for multi-document Arabic summarisation. In: Salem M, Shaalan K, Oroumchian F, Shakery A, Khelalfa H (eds) The 7th Asian Information Retrieval Societies (AIRS 2011), Springer Berlin / Heidelberg, Lecture Notes in Computer Science, vol 7097, pp 550–561

El-Haj M, Kruschwitz U, Fox C (2011b) Multi-document Arabic text summarisation. In: The 3rd Computer Science and Electronic Engineering Conference (CEEC'11), IEEE Xplore, Colchester, UK

El-Haj M, Kruschwitz U, Fox C (2011c) University of Essex at the TAC 2011 multilingual summarisation pilot. In: Text Analysis Conference (TAC) 2011, MultiLing Summarisation Pilot, TAC, Maryland, USA

Fattah M, Ren F (2008) Automatic text summarization. In: Proceedings of World Academy of Science, World Academy of Science, vol 27, pp 192–195

Fellbaum C (ed) (1998) WordNet: An Electronic Lexical Database (Language, Speech, and Communication), illustrated edition edn. The MIT Press

Foster I, Kesselman C, Nick J, Tuecke S (2002) Grid services for distributed system integration. Computer 35(6):37–46

Fukumoto F, Sakai A, Suzuki Y (2010) Eliminating redundancy by spectral relaxation for multi-document summarization. In: Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, TextGraphs-5, pp 98–

102
Getao K, Miriti E (2006) Automatic construction of a kiswahili corpus from the world wide web. In: Measuring Computing Research Excellence and Vitality, pp 209–219
Giannakopoulos G, Karkaletsis V (2011) AutoSummENG and MeMoG in evaluating guided summaries. In: The Proceedings of the Text Analysis Conference, TAC, MD, USA
Giannakopoulos G, Karkaletsis V, Vouros G, Stamatopoulos P (2008) Summarization system evaluation revisited: N–gram graphs. ACM Transactions on Speech and Language Processing (TSLP) 5(3):1–39
Giannakopoulos G, El-Haj M, Favre B, Litvak M, Steinberger J, Varma V (2011) TAC 2011 multiling pilot overview. In: Text Analysis Conference (TAC) 2011, MultiLing Summarisation Pilot, TAC, Maryland, USA
Graff D (2003) Arabic Gigaword. Linguistic Data Consortium, Philadelphia, lDC catalogue number: LDC2003T12, ISBN: 1–58563–271–6
Graff D, Chen K, Kong J, Maeda K (2006) Arabic Gigaword second edition. Linguistic Data Consortium, Philadelphia, lDC catalogue number: LDC2006T02, ISBN: 1–58563–371–2
Green N, Larasati S, Žabokrtský Z (2012) Indonesian dependency treebank: Annotation and parsing. In: Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, Faculty of Computer Science, Universitas Indonesia, Bali,Indonesia, pp 137–145, URL http://www.aclweb.org/anthology/Y12-1014
Guevara E (2010) Nowac: a large web-based corpus for norwegian. In: Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, WAC-6 '10, pp 1–7
Habash N, Roth R (2011) Using deep morphology to improve automatic error detection in Arabic handwriting recognition. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pp 875–884
Haddad B, Yaseen M (2005) A compositional approach towards semantic representation and construction of ARABIC. In: Proceedings of the 5th International Conference on Logical Aspects of Computational Linguistics, Springer-Verlag, Berlin, Heidelberg, LACL'05, pp 147–161
Hajic J, Smrz O, Zemanek P, Pajas P, Snaidauf J, Beska E, Kracmar J, Hassanova K (2004) Prague Arabic dependency treebank 1.0. Linguistic Data Consortium, Philadelphia, lDC catalogue number: LDC2004T23, ISBN: 1–58563–319–4
Halpern J (2006) The contribution of lexical resources to natural language processing of cjk languages. In: Huo Q, Ma B, Chng ES, Li H (eds) Chinese Spoken Language Processing, Lecture Notes in Computer Science, vol 4274, Springer Berlin Heidelberg, pp 768–780, DOI 10.1007/11939993_77, URL http://dx.doi.org/10.1007/11939993_77
Hendrickx I, Daelemans W, Marsi E, Krahmer E (2009) Reducing redundancy in multi-document summarization using lexical semantic similarity. In: Proceedings of the 2009 Workshop on Language Generation and Summarisation, Association for Computational Linguistics, Stroudsburg, PA, USA, UCNLG+Sum '09, pp 63–66
Hmeidi I, Al-Shalabi R, Al-Taani A, Najadat H, Al-Hazaimeh S (2010) A novel approach to the extraction of roots from Arabic words using bigrams. Journal of the American Society for Information Science and Technology 61(3):583–591
Howe J (2008) Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Publishing Group
Huang S, Graff D, Doddington G (2002) Multiple-translation chinese corpus. Linguistic Data Consortium, Philadelphia, lDC catalogue number: LDC2002T01, ISBN: 1–58563–217–1
Jing H, McKeown K (1998) Combining multiple, large-scale resources in a reusable lexicon for natural language generation. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics — Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '98, pp 607–613, DOI 10.3115/980845.980946, URL http://dx.doi.org/10.3115/980845.980946
Kaisser M, Lowe J (2008) Creating a research collection of question answer sentence pairs with amazon's mechanical turk. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC, Marrakech, Morocco
Katragadda R, Pingali P, Varma V (2009) Sentence position revisited: A robust light–weight update summarization 'baseline' algorithm. In: Proceedings of the Third International Workshop on Cross Lingual Information Access CLIAWS3'09, Association for Computational Linguistics, Morristown, NJ, USA, pp 46–52

Kazai G, Kamps J, Koolen M, Milic-Frayling N (2011) Crowdsourcing for book search evaluation: Impact of Hit design on comparative system ranking. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '11, pp 205–214

Kilgarriff A, Charalabopoulou F, Gavrilidou M, Johannessen J, Khalil S, Johansson Kokkinakis S, Lew R, Sharoff S, Vadlapudi R, Volodina E (2013) Corpus-based vocabulary lists for language learners for nine languages. Language Resources and Evaluation (LREV) pp 1–43, DOI 10.1007/s10579-013-9251-2, URL http://dx.doi.org/10.1007/s10579-013-9251-2

Kittur A, Smus B, Khamkar S, Kraut R (2011) CrowdForge: Crowdsourcing complex work. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, ACM, New York, NY, USA, UIST '11, pp 43–52

Kozareva Z, Hovy E (2013) Tailoring the automated construction of large-scale taxonomies using the web. Language Resources and Evaluation (LREV) 47(3):859–890, DOI 10.1007/s10579-013-9229-0, URL http://dx.doi.org/10.1007/s10579-013-9229-0

Larkey L, Ballesteros L, Connell M (2002) Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '02, pp 275–282

Li P, Zhu Q, Qian P, Fox G (2007) Constructing a large scale text corpus based on the grid and trustworthiness. In: Proceedings of the 10th International Conference on Text, Speech and Dialogue, Springer-Verlag, Berlin, Heidelberg, TSD'07, pp 56–65

Lin C (2004) ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), WAS 2004), pp 25–26

Lloret E, Plaza L, Aker A (2013) Analyzing the capabilities of crowdsourcing services for text summarization. Language Resources and Evaluation (LREV) 47(2):337–369

Luhn H (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development 2(2):159–165

Maamouri M, Bies A, Jin H, Buckwalter T (2003) Arabic treebank: Part 1 v 2.0. Linguistic Data Consortium, Philadelphia, lDC catalogue number: LDC2003T06, ISBN: 1–58563–261–9

Maamouri M, Bies A, Buckwalter T, Jin H (2004) Arabic treebank: Part 2 v 2.0. Linguistic Data Consortium, Philadelphia, lDC catalogue number: LDC2004T02, ISBN: 1–58563–282–1

Maamouri M, Bies A, Buckwalter T, Jin H, Mekki W (2005) Arabic treebank: Part 3 (full corpus) v 2.0 (MPG + syntactic analysis). Linguistic Data Consortium, Philadelphia, lDC Catalogue number: LDC2005T20, ISBN: 1–58563–341–0

Maegaard B, Atiyya M, Choukri K, Krauwer S, Mokbel C, Yaseen M (2008) Medar: Collaboration between European and Mediterranean Arabic partners to support the development of language technology for Arabic. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC, Marrakech, Morocco

Marcus M, Marcinkiewicz M, Santorini B (1993) Building a large annotated corpus of english: The penn treebank. Computational Linguistics 19(2):313–330, URL http://dl.acm.org/citation.cfm?id=972470.972475

Marsi E, Krahmer E (2013) Construction of an aligned monolingual treebank for studying semantic similarity. Language Resources and Evaluation (LREV) pp 1–28, DOI 10.1007/s10579-013-9252-1, URL http://dx.doi.org/10.1007/s10579-013-9252-1

zu Meyer S, Stein B, Kulig M (2007) Plagiarism detection without reference collections. In: Decker R, Lenz H (eds) Advancesin Data Analysis (Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8–10, 2006), Springer Berlin Heidelberg, Studies in Classification, Data Analysis, and Knowledge Organization, pp 359–366

Mourad A, Darwish K (2013) Subjectivity and sentiment analysis of Modern Standard Arabic and Arabic microblogs. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Atlanta, Georgia, pp 55–64, URL http://www.aclweb.org/anthology/W13-1608

Nemeskey D, Simon E (2012) Automatically generated ne tagged corpora for english and hungarian. In: Proceedings of the 4th Named Entity Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, NEWS '12, pp 38–46

Nenkova A (2005) Automatic text summarization of newswire: Lessons learned from the document understanding conference. In: Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3, AAAI Press, AAAI'05, pp 1436–1441

Nenkova A, McKeown K (2012) A survey of text summarization techniques. In: Aggarwal CC, Zhai C (eds) Mining Text Data, Springer US, pp 43–76, DOI 10.1007/978-1-4614-3223-4_3

Nganga W (2012) Building Swahili resource grammars for the grammatical framework. In: Shall We Play the Festschrift Game?, Springer Berlin Heidelberg

Nguyen P, Vu X, Nguyen T, Nguyen V, Le H (2009) Building a large syntactically-annotated corpus of vietnamese. In: Proceedings of the Third Linguistic Annotation Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL-IJCNLP '09, pp 182–185

no ABC, Rosso P, Agirre E, Labaka G (2010) Plagiarism detection across distant language pairs. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '10, pp 37–45

Outahajala M, Benajiba Y, Rosso P, Zenkouar L (2011) Pos tagging in Amazighe using support vector machines and conditional random fields. In: Natural Language Processing and Information Systems, Springer Berlin Heidelberg, pp 238–241

Poesio M, Chamberlain J, Kruschwitz U, Robaldo L, Ducceschi L (2013) Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. ACM Trans Interact Intell Syst 3(1):3:1–3:44

Potthast M, Hagen M, Gollub T, Tippmann M, Kiesel J, Rosso P, Stamatatos E, Stein B (2013) Overview of the 5th international competition on plagiarism detection. In: CLEF 2013 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN-13), Valencia, Spain, pp 1–30

Prochazka S (2006) "Arabic" Encyclopedia of Language and Linguistics, vol 1, 2nd edn. Elsevier

Ptaszynski M, Rzepka R, Araki K, Momouchi Y (2012) Automatically annotating a five-billion-word corpus of japanese blogs for affect and sentiment analysis. In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics, Stroudsburg, PA, USA, WASSA '12, pp 89–98

Radev D, Jing H, Budzikowska M (2000) Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization — Volume 4, Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL-ANLP-AutoSum '00, pp 21–30

Radev D, Jing H, Sty M, Tam D (2004) Centroid-based summarization of multiple documents. Information Processing and Management 40:919–938, DOI 10.1016/j.ipm.2003.10.006

Roberts A, Al-Sulaiti L, Atwell E (2006) aConCorde: Towards an open-source, extendable concordancer for Arabic. Corpora 1(1):39–60, DOI 10.3366/cor.2006.1.1.39

Sarkar K (2009) Centroid-based summarization of multiple documents. TECHNIA – International Journal of Computing Science and Communication Technologies 2

Sawalha M, Atwell E (2010a) Constructing and using broad-coverage lexical resource for enhancing morphological analysis of Arabic. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) The 7th Language Resources and Evaluation Conference LREC, LREC 2010, Valletta, Malta, pp 282–287

Sawalha M, Atwell E (2010b) Fine-grain morphological analyzer and part-of-speech tagger for Arabic text. In: The 7th Language Resources and Evaluation Conference LREC, LREC 2010, Valletta, Malta, pp 1258–1265

Schalley A (2012) Ontology and the lexicon: a natural language processing perspective. (studies in natural language processing.). Language Resources and Evaluation (LREV) 46(1):95–100, DOI 10.1007/s10579-011-9138-z, URL http://dx.doi.org/10.1007/s10579-011-9138-z

Sekine S, Nobata C (2003) A survey for multi-document summarization. In: Proceedings of the HLT-NAACL 03 on Text Summarization Workshop - Volume 5, Association for Computational Linguistics, Stroudsburg, PA, USA, HLT-NAACL-DUC '03, pp 65–72

Smrž O (2007) ElixirFM: Implementation of functional Arabic morphology. In: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Association for Computational Linguistics, Stroudsburg, PA, USA, Semitic '07, pp 1–8

Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp 254–263

Walther G, Sagot B (2010) Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish. In: Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop), Valetta, Malta

Wang D, Li T (2012) Weighted consensus multi-document summarization. Information Processing and Management 48(3):513–523

Wang S, Li W, Wang F, Deng H (2010) A survey on automatic summarization. In: Information Technology and Applications (IFITA), 2010 International Forum on, vol 1, pp 193–196

Wilks Y, Fass D, Guo C, McDonald J, Plate T, Slator B (1988) Machine tractable dictionaries as tools and resources for natural language processing. In: Proceedings of the 12th conference on Computational linguistics — Volume 2, Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '88, pp 750–755, DOI 10.3115/991719.991789, URL http://dx.doi.org/10.3115/991719.991789

Yang Y, Bansal N, Dakka W, Ipeirotis P, Koudas N, Papadias D (2009) Query by document. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA, WSDM '09, pp 34–43

Yaseen M, Theophilopoulos N (2001) NAPLUS: Natural Arabic processing for language understanding systems

Yeh J, Ke H, Yang W (2008) iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. Expert Systems with Applications 35(3):1451–1462

Appendix

A EASC Corpus Guidelines Appendix

Creating EASC Corpus Guidelines

The Mechanical Turk workers were given the following guidelines for completing the task of creating single-document summaries corpus (Section 3).

1. Read the Arabic sentences (Document).
2. In the first text box below type the number of sentences that you think are focusing on the main idea of the document.
3. The number of the selected sentences should not exceed 50% of the articles‘ sentences, for example if there are 5 sentences you can only chose up to 2 sentences.
4. Just add the numbers of the sentences, please do not write text in the first text box. For example (1,2,4).
5. In the second text box write down one to three Keyword(s) that represent the main idea of the documents, or keywords that can be used as title for the article, do not exceed 3 keywords.
6. If you have any comments please add them to the third text box below.
7. Failing to follow the guides correctly could lead to task rejection
8. NOTE: The article chosen for this task was selected randomly from the Internet. The purpose of this task is purely educational and does not reflect, support or contradict with any opinion or point of view.

Figure 5, shows an example for one of the hits provided to the workers on Mechanical Turk website. The reason for the selection of an answer field (instead of a checkbox or radio button) is that we aimed to reduce the noise and track spammers. We think that a design with e.g. radio buttons is not able to distinguish between MTurks as spammers and MTurks who produce noise (wrong selection but not produced by random procedure as it is the case by spammers), for example if the worker wrote “two” instead of “2”, we still consider this as a valid answer. Using radio buttons, when selection is made it is not clear whether the MTurk worker has selected it because he thought it is a correct answer or just by random. However, if we force an MTurk to write down the answer then this gives us the possibility to distinguish between spammers and noise. A spammer would give answers which are composed by random characters and/or numbers whereas a noise could be close to the right answer.

B TAC-2011 Dataset Guidelines Appendix

Creating TAC-2011 Dataset Guidelines

The following task guidelines were required by the participants to create a manual corpus for TAC-2011 MultiLing Pilot:

1. **Translation:** Given the source language text A, the translator is requested to translate each sentence in A, into the target language. Each target sentence should keep the meaning from the source language. The resulting text would be a UTF8 encoded plain text file, named $A.[lang]$, where $[lang]$ should be replaced by the target language. For each text the following check list should be followed:
   - The translator notes down the starting time for the reading step.
   - The translator reads the source text at least once to get an understanding.
   - The translator notes down the starting time for the translation step.
   - Perform the translation.
   - The translator notes down the finishing time for the translation step.
2. **Translation Validation:** After the completion of each translation, another translator “validator” should verify the correctness of the output. If errors are found, then the validator is to perform any corrections and finalise the translation. For each text the following check list should be followed:
   - The translator notes down the starting time for the verification step.
   - Read the translation and verify the text. Perform any corrections needed.
   - The translator notes down the finishing time for the verification step.

Fig. 5 EASC: MTurk Hit Example

3. **Summarisation:** The summariser will read the whole set of texts at least once. Then, the summariser should compose a summary, with a minimum size of 240 and a maximum size of 250 words. The summary should be in the same language as the texts in the set. The aim is to create a summary that covers all the major points of the document set (what is major is left to summariser discretion). The summary should be written using fluent, easily readable language. No formatting or other markup should be included in the text. The output summary should be a self-sufficient, clearly written text, providing no other information than what is included in the source documents. For each document set the following check list should be followed:
   - The summariser notes down the starting time for the reading step.
   - Read the whole set of texts in the document set at least once, to have an overall understanding of the event(s) described.
   - The summariser notes down the starting time for the summarisation step.
   - The summariser writes the summary, reviewing the source texts, if required.
   - The summariser notes down the end time for the summarisation step.
4. **Evaluation:** Each summary will be graded by 3 evaluators. If the summarisers are used as evaluators, no self-evaluation should be allowed. Evaluators read each translated document set at least once. Then they read the summary they are to evaluate, and they grade it. 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.

C Evaluating Arabic Summaries Guidelines

Evaluating EASC Corpus Summaries Guidelines

1. Read a document at least once.
2. In the Evaluation form note the time (in minutes) it took you to read the document. Figure 6, shows a sample of the form provided to the evaluators.
3. Read the document's summary you are to evaluate
4. In the Evaluation form grade the summary with a value between 1 and 5.
   - Each summary is to be assigned an integer grade from 1 to 5, related to the overall responsiveness of the summary.
   - A summary is worth a 5, if it appears to cover all the important aspects of the corresponding document using fluent, readable language.
   - A summary is worth 1, if it is either unreadable, nonsensical, or contains only trivial information from the document.
   - We consider the content and the quality of the language to be equally important in the grading.

Evaluator Name	Mahmoud El-Haj

Document	Summary	Score
01.txt	0101.txt	3
01.txt	0201.txt	3
01.txt	0301.txt	3
01.txt	0401.txt	3
01.txt	0501.txt	3
01.txt	0701.txt	3
01.txt	0801.txt	3
01.txt	0601A.txt	3
01.txt	0601B.txt	3
01.txt	0601C.txt	3

Document	Topic	Reading Time (minutes)
01.txt	البترول	10
02.txt	الإقتصاد الخليجي	10
03.txt	الإثنا عشري	10
04.txt	مقابلة ملك الأردن	10
05.txt	إيطاليا	10
06.txt	المسيحية	10
07.txt	شاشات اللمس	10
08.txt	الصاروخ	10
09.txt	الأردن	10
10.txt	الأهرامات	10

Fig. 6 Human Experts Evaluation Form Sample