UNIVERSITY OF LEEDS

This is a repository copy of *Accelerating the processing of large corpora: using Grid Computing for lemmatizing the 176 million words Arabic Internet Corpus*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/81622/

Proceedings Paper:
Sawalha, M and Atwell, ES (2013) Accelerating the processing of large corpora: using Grid Computing for lemmatizing the 176 million words Arabic Internet Corpus. In: Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2. The 2nd Workshop of Arabic Corpus Linguistics WACL-2, 22-26 Jul 2013, Lancaster University, UK. UCREL .

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Accelerating the Processing of Large Corpora: Using Grid Computing Technologies for Lemmatizing 176 Million Words Arabic Internet Corpus

Majdi Sawalha and Eric Atwell
University of Leeds, Leeds, LS2 9JT, UK
mailto:sawalha@comp.leeds.ac.uk, eric@comp.leeds.ac.uk

The Arabic Internet Corpus is one of several large corpora collected for Translation Studies research at http://corpus.leeds.ac.uk/internet.html alongside Internet Corpora of English, Chinese, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian and Spanish (Sharoff, 2006). The Arabic Internet Corpus consists of about 176 million words[1]. Initially it consisted of raw text, with no further processing such as lemmatization or part-of-speech tagging. In this paper we show how we added the lemma and root for each word.

Arabic is a morphologically rich and highly inflectional language. Hundreds of words can be derived from the same root; and a lemma can appear in the text in many different forms due to the glutination of clitics at the front and end of the word. Therefore, lemmatization and root extraction is necessary for search applications, to enable inflected forms of a word to be grouped together. We used the lemmatizing part of an Arabic morphological analyzer (Sawalha and Atwell, 2009, Sawalha and Atwell, 2010) to annotate the Arabic Internet Corpus words at two levels; the lemma and the root, illustrated in Figure 1. The morphological analyzer is relatively slow. In initial tests it processed 7 words per second, because the analyzer has to deal with orthographic issues, spell checking of the word's letters, short vowels and diacritics and the large dictionaries provided to the analyzer. An estimate execution time for lemmatizing the full Arabic Internet Corpus was 300 days using ordinary uni-processor machine.

To reduce the processing time of the whole task, we used the power of HPC (High Performance Computing). NGS[2] (National Grid Services) aims to enable coherent electronic access for UK researchers to all computational and data based resources and facilities required to carry out their research, independent of resource or researcher location. We used the huge computational power of NGS to lemmatize the Arabic internet corpus and we gained massive reduction in execution time. We divided the Arabic Web Corpus into half-million-word files. Then we wrote a program that generates scripts to run the lemmatizer for each file in parallel. The output files are combined in one lemmatized Arabic Internet Corpus, comprising 176 million word-tokens, 2,412,983 word-types, 322,464 lemma-types, and 87,068 root-types.

By using the NGS we massively reduced the execution time of processing the 176M-word corpus to only 5 days. It might have been a few hours, had we been able to allocate enough CPUs to process all files strictly in parallel; NGS provides virtual parallel processing on a reduced set of CPUs. After the output files were combined into one lemmatized Arabic Web Corpus, 10 random samples, of 100 words each, were selected to evaluate the accuracy of the lemmatizer. For each sample, we computed the accuracy of

[1] The frequency list of the Arabic internet corpus http://corpus.leeds.ac.uk/frqc/i-ar-forms.num
[2] NGS (National Grid Services) http://www.ngs.ac.uk

the root and lemma analysis.. We found that the average root and lemma accuracy was consistent across samples. The average root accuracy was about 81.20% and the average lemma accuracy was 80.80%; see Figure 2.

	طول	طويل	طويلا				علل	لعله
	،	،	،	STOP_WORD	أن	أن	أن	أن
	جلس	جلس	وجلست	STOP_WORD	كون	كان	كون	يكون
N_BP	بيت	بيت	البيوت		كبس	كابوس	كبس	كابوسا
	سكن	ساكن	ساكنة		فوق	يستفيق	فوق	ويستفيق
	،	،	،	STOP_WORD	منه	منه	منه	منه
	طرق	مطرق	مطرقة	STOP_WORD	على	على	على	على
	،	،	،		شيأ	أشياء	شيأ	الأشياء
	صبح	مصابيح	والمصابيح		ءلف	أليف	ءلف	الأليفة
	صفر	صفراء	الصفراء		طيب	طيب	طيب	والطيبة
	قرر	مقرور	المقرورة		حبب	حبيب	حبب	والحبيبة
	زفف	نزف	تترف		.	.	.	.
	ضوأ	ضوء	ضوءا		مدد	امتد	مدد	وامتد
					شرع	شارع	شرع	الشارع
					ضيق	ضيق	ضيق	الضيق
					طول	طويل	طول	طويلا

Figure 1: Sample of lemmatized sentence from the Arabic Internet Corpus, لعله أن يكون كابوسا ويستفيق منه على الأشياء الأليفة والطيبة والحبيبة. وامتد الشارع الضيق طويلا.. طويلا، وجلست البيوت ساكنة، مطرقة، والمصابيح الصفراء المقرورة تترف ضوءا. *la'alhu 'an yakūna kābūs^{an} wa yastafīqu minhu 'alā al-'šyā'i al-'alīfa^{ti} wa aṭ-ṭyyeba^{ti} wa al-ḥabība^{ti}. wa imtadda aš-šāri'u al-ḍayyiqu ṭawīl^{an}.. ṭawīl^{an} wa ğalasat al-buyūtu sākinat^{un}, muṭriqat^{un}, wa al-maṣābīḥu aṣ-ṣafrā'u al-maqrūra^{tu} tanzifu ḍū'^{an}*
Perhaps it is a nightmare and he will wake up to the usual, good and beloved things. The narrow road is extended long.. Long, the home sat silent, listening, speechless, and the yellow bubbled lamps bleeding light.

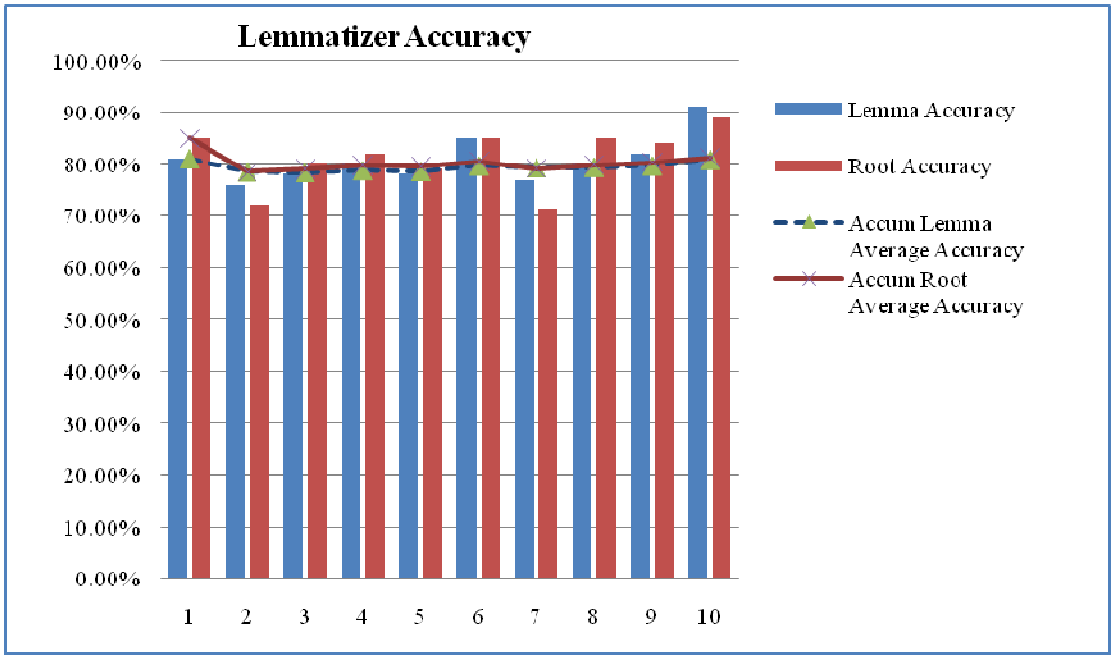

Figure 2: Lemma and root accuracy of the lemmatized Arabic internet corpus

References

SAWALHA, M. & ATWELL, E. (2009) Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. *Proceedings of the 5th International Corpus Linguuistics Conference CL2009.* Liverpool, UK.

SAWALHA, M. & ATWELL, E. (2010) Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. *Language Resource and Evaluation Conference LREC 2010* Valleta, Malta.

SHAROFF, S. (2006) Creating General-Purpose Corpus Using Automated Search Engine Queries. IN BARONI, M. & BERNARDINI, S. (Eds.) *WaCky! Working papers on the Web as Corpus.* Bologna, GEDIT.