Abs-49

Majdi Sawalha and Eric Atwell (both University of Leeds)

Accelerating the processing of large corpora: Using grid computing technologies for lemmatizing the 176 million word Arabic Internet Corpus

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. 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 beginning and at the 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 (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 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.

لعله	عل	علل		طويلا	طويل	طول
أن	أن	أن	STOP_WORD	.	.	.
يكون	كان	كون	STOP_WORD	.	.	.

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

Figure 1: Sample of lemmatized sentence from the Arabic Internet Corpus

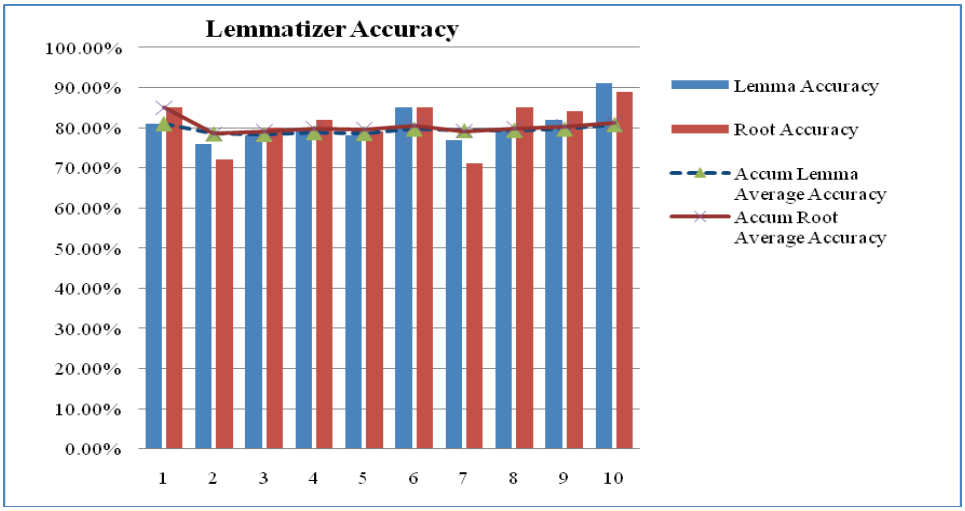

Figure 2: Lemma and root accuracy of the lemmatized Arabic internet corpus

References

Sawalha, Majdi; Atwell, Eric (2009). *Linguistically Informed and Corpus Informed Morphological Analysis of Arabic*. in: **Proceedings of the 5th International Corpus Linguistics Conference CL2009**, 20-23 July 2009, Liverpool, UK.

Sawalha, Majdi; Atwell, Eric (2010). *Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text.* in: **Proceedings of the Language Resource and Evaluation Conference LREC 2010**, 17-23 May 2010, Valleta, Malta.

Sharoff, Serge (2006). Creating General-Purpose Corpus Using Automated Search Engine Queries. In M. Baroni and S. Bernardini (eds.). *WaCky! Working papers on the Web as Corpus*, pp. 63-98. Bologna: GEDIT.