Two approaches to gathering text corpora from the World Wide Web

Gerrit Botha, Etienne Barnard

Human Language Technologies Research Group,
University of Pretoria / Meraka Institute, Pretoria, South Africa
gbotha@meraka.co.za, ebarnard@up.ac.za

Abstract

Many applications of pattern recognition to natural language processing require large text corpora in a specified language. For many of the languages of the world, such corpora are not readily available, but significant quantities of text are available on the World Wide Web. We describe and compare two approaches to gathering language-specific corpora from this resource, and show that the use of a commercial search engine as a first stage leads to good results.

1. Introduction

Many of the most promising applications of pattern recognition involve human language: for example, word statistics, information retrieval, speech recognition, topic classification, and machine translation all involve the application of principles from pattern recognition to human languages[1][2][3]. Since the statistical approach to pattern recognition depends greatly on the availability of sufficient training data, it is crucial that significant language-oriented databases (known as "corpora") should be collected[4][5]. The most important corpora in practice are collections of speech data and of text data; the latter category is the focus of the current paper.

Text corpora are used for a variety of purposes, ranging from the computation of word and letter statistics to the training of part-of-speech taggers and translation models. In light of the variability of language, very large corpora are required for the more sophisticated applications; for example, the Chinese Gigaword Corpora, the Multilingual News Text and the Arabic Treebank[6]. Although such large corpora exist for most of the "large" languages of the world (Mandarin Chinese, Hindi, English, French, etc.), it is difficult to collect enough text for the vast majority of languages. Even when copious text is available in book format, difficulties such as copyright issues and the manual labour required to scan these books in mean that no corpora are generally available for most indigenous languages in the world.

Against this background, the Internet is an extremely valuable resource. Very large amounts of text are available on the Internet in electronic format, much of it not copyrighted[7][8]. The diversity of languages on the Internet is also significant, and even in relatively "small" languages, significant numbers of articles are available on the World Wide Web (WWW). However, these articles are widely scattered, and it is generally a difficult and time-consuming task to gather significant amounts of text in a specified language from the WWW.

We therefore investigate the automation of language-specific text collection on the WWW. In particular, we study two methods for performing this task – one based on Web crawlers (Section 2), and the other using search engines (Section 3). Comparative results and a discussion of future directions are presented in Section 4.

2. Web crawlers for text collection

A Web crawler or robot is a program that recursively visits web pages to collect information for a specific need. This allows a user to collect information without significant human interaction. A good example is Googlebot, Google's web crawler, which collects documents from the web to build a searchable index for the Google search engine. The basic implementation of the crawler is as follows: the crawler starts at a root page and then follows the links on that page (see Fig. 1). The contents of the returned pages are processed, and the links on each of these pages are subsequently followed. The crawling process continues in this tree-like manner, collecting content as it goes on. Sophisticated algorithms to make crawlers more efficient have been developed[9]. For the purpose of experimenting with multilingual text corpora, we implemented a basic crawler and studied the results obtained – the philosophy being that a more robust and efficient implementation could be developed if good initial results were found.

Our basic idea was to crawl a large domain on the Internet and find text that contains pre-specified words from a language. Text found in this way is likely to belong to the desired language. The text can be saved in a database and the words in the text can now be used as additional keys to finding more text. Once a sufficient corpus of documents has been found, an n-gram language model[10][11] can be trained on the valid documents, and used to select additional valid documents in the target language.

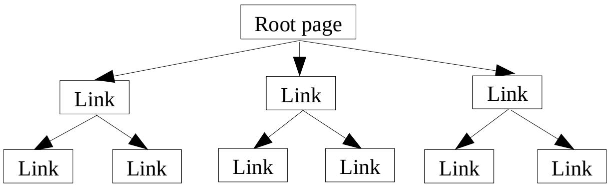

Figure 1: Basic architecture of a Web crawler: starting with an initial document, the system recursively follows all the links in all subsequent documents.

We started crawling form a website that contains only text in Zulu, hoping this would give a good root for our crawler (which was implemented as in Fig. 1). The crawler was started with a basic set of 300 words in Zulu, and crawled the Web for a two-day period. The results were disappointing, as shown in Fig. 2: 45,000 words out of a total of 10,565,000 words belonged to Zulu or consider it as 200Kb out of a total of 24Mb contained Zulu text. The text were mostly collected from the website we used as starting point. The basic conclusion was that there isn't really a large domain of interconnected sites that contain only Zulu websites with relevant content. The overwhelming majority of Web sites visited in this way are in fact predominantly or exclusively in English, and a large database of English text can thus be created – but this process was of little use for the creation of a Zulu text corpus. In fact, we did not retrieve sufficiently many Zulu documents in this fashion to build a robust n-gram model.

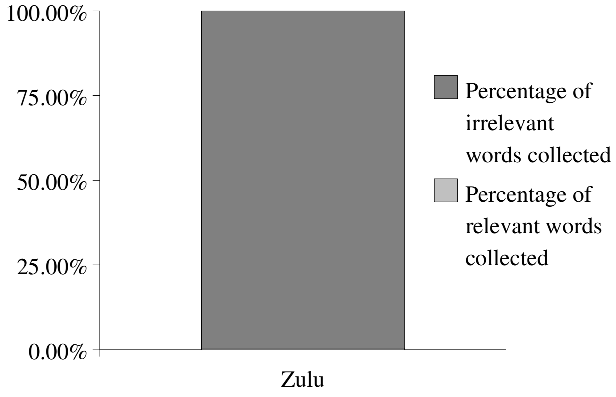

Figure 2: Search results when using the basic crawler architecture to search for Zulu text: the vast majority of crawled documents contain only English text, and limited additional Zulu data is found.

3. Using a search engine for text collection

In order to address the preponderance of English documents found with our basic Web crawler, a more directed approach was developed, as shown in Fig. 3. An Internet search engine limited to the geographic region most likely to contain documents in the target language is used to find multiple initial documents for the Web crawler. (For the case of Zulu, we used the commercial search engine Google [12], and limited the search to South African domains.) A series of words in the target language are used as search terms in the selected engine, and the crawler is modified to harvest the documents returned by the search engine. The language of the document were then identified by using n-gram language models of South African languages[1].

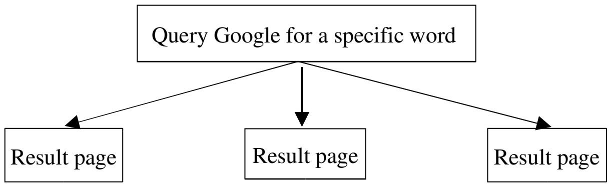

Figure 3: A more directed approach, by using the Google search engine to search for relevant text.

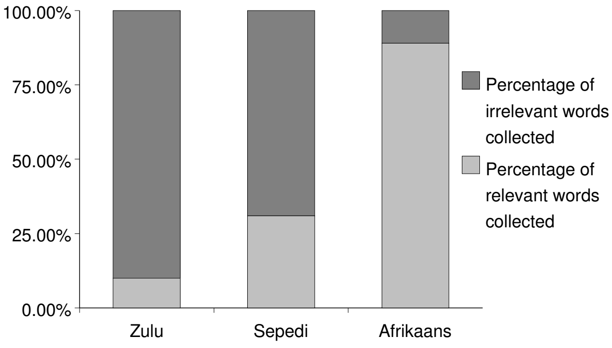

Figure 4: Search results when using the search-based architecture to search for Zulu, Sepedi and Afrikaans text: the focused nature of the search ensured that more usable text was collected from a smaller set of visited sites.

This approach was tested for three South African languages, namely Afrikaans, Sepedi (SeSotho sa Laboa) and Zulu. Again, 300 words from each target language were used as starting point for the Web search. As shown in Fig. 4, a significant corpus of text was collected in this fashion[2]. Although fewer Web sites were visited, much

[1] Identification can be improved by using the collected text to build a larger language model, or by using more sophisticated classification – for example, based on Transformation Based Learning [13]. We have not performed these refinements, since our baseline system functioned with acceptable accuracy in spot tests.

[2] Zulu and Ndebele are highly similar in written form, and discrimination within this pair is difficult. Given the preponderance of Zulu data on the Web, the majority of our documents are in fact in Zulu, but the presence of a few Ndebele documents renders our estimate approximate.

more text was obtained. In detail, the following results were obtained: 228,000 words out of a total of 2,280,000 words belonged to Zulu (1.14Mb out of a total of 11.4Mb contained relevant Zulu text). 462,000 words out of a total of 1,540,000 words belonged to Sepedi (2.4Mb out of a total of 7.4Mb). 10,893,600 words out of a total of 12,240,000 words belonged to Afrikaans (28.8Mb out of a total of 30Mb).

4. Conclusion

A basic Web-crawling approach to the collection of text corpora in smaller languages does not succeed, in light of the preponderance of world-language documents retrieved from even a carefully-selected starting point. It is therefore preferable to pre-filter the search process by selecting a multiplicity of sites which all contain at least one word in the target language. This approach yielded highly promising results in three languages that were evaluated in our research.

It remains to be seen whether a hybrid approach can be constructed to further expand the set of sites obtained. A Web crawler can, for example, be initiated from each of the sites returned by an initial search, and the language model trained from those sites used to filter other sited found during recursive traversal of the Web. Since the proportion of valid target-language documents is likely to be very small (based on the results in Section 2), this approach must be highly efficient. We are currently investigating such a methodology.

5. References

[1] D. Jurafsky, J.H. Martin, "Speech and Language Processing", 2nd ed., Pearson Education Publishers, pp 799–827.

[2] T. Matsuoka, R. Hasson, M. Barlow, S. Furui, "Language Acuisition From A Text Corpus For Speech UnderStanding", ICASSP-96. Conference Proceedings, 1996.

[3] G. Dalkilic, Y. Cebi, "Word Statistics of Turkish Language on a Large Scale Text Corpus", ITCC International Conference, Volume 2, pp 319-324, 2004.

[4] N.S Dash, B.B. Chaudhuri, "Using Text Corpora for Understanding Polysemy in Bangla", Language Engineering Conference, pp 99 - 109, December 13-15, 2002.

[5] M. Xiaoyi, "Text collection at Linguistic Data Consortium", Machine Translation Summit VII, Kent Ridge Digital Labs, National University of Singapore September 16, 1999.

[6] Linguistic Data Consortium [Online]. Available: http://www.ldc.upenn.edu

[7] M. Xiaoyi, M. Liberman, "A Method for Bilingual Search over the Internet", Machine Translation Summit VII, Kent Ridge Digital Labs, National University of Singapore, September 13, 1999.

[8] O. Salor, B. Pellom, T. Ciloglu, K. Hacioglu, M. Demirekler, "On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language", ICSLP-2002, International Conference on Spoken Language Processing, Denver Colorado,USA, September 16-20, 2002.

[9] J. Qin, H. Chen, "Using Genetic Algorithm in Building Domain-Specific Collections: An Experiment in the Nanothechnology Domain", Proceedings of the 38th Hawaii International Conference on System Sience, 2005.

[10] D. Jurafsky, J.H. Martin, "Speech and Language Processing", Pearson Education Publishers, pp 287-319.

[11] X. Huang, A. Acero, H. Hon, "Spoken Language Processing", UK: Prentice Hall Publishers, pp 558-584.

[12] Google [Online]. Available: http://www.google.co.za

[13] J.C. Marcadet, V. FIscher, C. Waast-Richard "A transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis", Interspeech, Lisbon, Portugal, pp. 2249-2252 , 2005.