# WEB BASED SENTENCE COLLECTOR

Erdinç UZUN, Yılmaz KILIÇASLAN, Erdem UÇAR
{erdincuzun, yilmazk, erdemu}@trakya.edu.tr

Trakya University, Engineering and Architecture Faculty,
Computer Engineering Department, Edirne-Turkey

## ABSTRACT

*The World Wide Web can be used as a source of machine-readable text for corpora. Search engines, programs that search documents for specified keywords and return a list of the documents, are the main tools by which such texts can be collected. However, the usefulness of results returned by search engines is limited at least by the sheer amount of noise on the Web. This study describes a Web Based Sentence Collector (WBSC) that uses search engines for retrieving Turkish documents and filters out any detected noise that degenerates the grammaticality of the sentences.*

*Keywords: World Wide Web, Corpora, Search Engine, Information Retrieval*

## INTRODUCTION

The one million word Brown corpus opened the chapter on computer-based language study in the early 1960s. Nowadays, the British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources. There are a lot of corpora prepared for English and other languages [American English Corpus (100 million words), Bank of English (525 million words), Helsinki Corpus (10 million words), Longman-Lancaster Corpus (14.5 million words), Oxford English Corpus (over 1 million words) and Scottish Corpus (4 million words)].

Available Turkish corpora are comparatively smaller: METU Turkish Corpus contains 2 million words [10] and Metu-Sabancı Turkish Treebank [13]. In certain cases, the amount of data might be required to be highly large. Manning and Schütze (1999) make the following observation:

> "In Statistical NLP, one commonly receives as a corpus a certain amount of data from a certain domain of interest, without having any say in how it is constructed. In such cases, having more training data is normally more useful than any concerns of balance, and one should simply use all the text that is available." (p. 20)

Under these circumstances, the Web may be considered as a corpus.

The World Wide Web, being essentially an enormous database of mostly textual documents, offers great opportunities to researchers. The web as a corpus entered this field with ACL meetings starting in 1999 [9, 11]. Kalgarriff and Grefenstetle (2001, 2003) proved the usefulness of web in this respect. They also underline the limitations of search engines [6, 7]. There are several attempts to overcome these limitations [1, 2, 4, and 14]. A number of tools have been developed to allow the web to be used a corpus: Gsearch system [3], WebCorp [5], KWiCFinder [4], the Linguist's Search Engine [7, 14].

The information explosion on the Web has placed high demands on search engines. Search engines often return thousands of documents in response to a user query. The search engines allow one to ask for content meeting specific criteria (containing a given word or phrase) and retrieve a list of items that match those criteria. This paper shows how to obtain Turkish grammatical sentences using search engines. For this purpose, we have developed a Web Based Sentence Collector (WBSC) that accepts a Turkish verb as input, then queries this verb using a search engine and returns grammatical Turkish sentences as output.

## WEB BASE SENTENCE COLLECTOR

This software has two main tasks:

- Generate a query for retrieving data from the search engine
- Parse the retrieved data and obtain grammatical Turkish sentences

A request to a search engine is a simple HTTP request. Search results are returned in either HTML or XML formats, in accordance with the search request. In this study, we have selected the HTML format.

WBSC, firstly, prepares a query string for an HTTP request. This query string contains one or more name-value pairs which are appropriate for search engines. The following is an example of query string for Google:

http://www.google.com.tr/search?hl=tr&q=gidecek&num=100&lr=tr

The query string starts with the host name or IP address, and is followed by the string "/search?" and one or more name-value pairs separated with an ampersand (&) character. When the query string is ready, WBSC sends an HTTP request. Then, the search engine produces results in HTML format.

The search engine results normally include a list of web pages with titles, a link to the page and a short description showing where the keywords match the content within a page. WBSC, secondly, parses these results and extracts a title, a link and a short description. Below is a pseudo code for parsing the retrieved data:

copy_text (S; Index, Count): Returns a substring of a string or a segment of a dynamic array.
S is an expression of a string or dynamic-array type. Index and Count are integer-type expressions. Copy returns a string containing a specified number of characters from a string or sub array containing Count elements starting at S[Index].
If Index is greater than the length of S, Copy returns a zero-length string ("") or an empty array.
If Count specifies more characters or array elements than are available, only the characters or elements from S[Index] to the end of S are returned.

add_sentence: adds appropriate sentences for Turkish.

```
x = 0  :pointer
temp_str: retrieved data from search engine results
start_link, start_title, start_description = False: True / False
text_title, text_description = ' ' : temporary variables

WHILE end of the temp_str

        IF copy_text (temp_str, x, 16) = '<a class=l href=' THEN
        x = x + 17;
        start_link := Doğru;
        END IF

        IF (basla_link = True) AND (temp_str[x]= '>') THEN
        x = x + 1;
        start_title = True;
        start_link = False;
        END IF

        IF (start_title = True)
           AND (copy_text (temp_str, x, 4)= '</a>') THEN
        x = x + 5;
        başla_title = False
        add_sentence(str_başlık)
        str_title = ' '
        END IF

        IF copy_text (temp_str, x, 26)= '<td class=j><font size=-1>' THEN
        x = x + 26;
        start_description = False
        END IF

        IF (start_description = True) AND
        (copy_text (temp_str, x, 18)= '<br><span class=a>') THEN
        x = x + 18;
        start_description = False
        add_sentence(str_description)
        str_description = ' ';
        END IF

IF start_title THEN
  str_title := str_title + temp_str[x];

IF start_description THEN
  str_description := str_description + temp_str[x];

x = x + 1;

END WHILE
```

Pseudocode 1. Parsing retrieved data

Then, WBSC starts a filtering process to obtain grammatical Turkish sentences from the title and short description. There are two filtering layers, the tasks of which are described in below:

- Remove HTML tags and illegal words and correct illegal characters.
- Eliminate similar sentences and sentences containing only one word.

Sentences may contain some problematic HTML tags (such as “<b>, </b>, <i>, </i>”) and illegal words (such as â€™ , &#39; , €œ , =&gt; , &quot). The first filtering layer removes these characters. Another task of this layer is to correct Turkish characters which are not supported by some software and web services (cf. Table 1). It is noteworthy that some software and web applications that have already been internationalized and successfully support many languages often lead to catastrophic failures when the language is Turkish.

| Illegal Character | In Turkish |
|---|---|
| Ã‡ , Ã§ | ç |
| Ä□ , ÄŸ | ğ |
| Ä° , Ã□ | i |
| Ä± , Ã½ | ı |
| Ã– , Ã¶ | ö |
| Å□ , ÅŸ | ş |
| Ãœ , Ã¼ | ü |

Table 1. Some Illegal Characters

The second filtering layer detects similar sentences and eliminates them. Furthermore, some sentences contain only one word. These sentences are also filtered out by WBSC.

## CONCLUSION

The search engines contain noisy information and misinformation. If these problems were removed, a search engine could be a very useful tool for language researchers. In this study, we describe a software module that searches a Turkish verb using search engines, then filters documents and returns grammatical sentences. In the future, we aim to use this software for automatically acquiring subcategorization frames and accomplishing word sense disambiguation for Turkish.

## REFERENCE

[1] Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In Proceedings of LREC 2004, 1313-1316.

[2] Baroni, M. and Sharoff, S. (2005). Creating special-ized and general corpora using automated search engine queries. Web as Corpus Workshop, Bir-mingham University, UK, 14th July 2005.

[3] Corley, S., Corley, M., Keller, F., Crocker, M., & Trewin, S. (2001). Finding Syntactic Structure in Unparsed Corpora: The Gsearch Corpus Query System. Computers and the Humanities, 35, 81-94.

[4] Fletcher, B. (2004). Making the Web more useful as a source for linguistic corpora. In Connor, U. and Upton, T. (eds.) Corpus linguistics in North America 2002, Amsterdam: Rodopi.

[5] Kehoe, A. and Renouf, A. (2002) WebCorp: Applying the Web to Linguistics and Linguistics to the Web. World Wide Web 2002 Conference, Honolulu, Ha-waii.

[6] Kilgarriff, A. (2001). The Web as corpus. Proceedings of Corpus Linguistics 2001.

[7] Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the Web as corpus. Computational Linguistics 29(3), 333-347.

[8] Manning, C. and Schütze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press, Cambridge

[9] Mihalcea, R. and Moldovan D. (1999). A method for word sense disambiguation of unrestricted text. In Proceedings of the 37th Meeting of ACL, pages 152–158, College Park, MD, June.

[10] Özge U. and Say B. (2004). Development of a Corpus Workbench for the METU Turkish Corpus. 4th Language Resources and Evaluation Conference, Lizbon, Portugal

[11] Resnik, P. (1999). Mining the Web for bilingual text. In Proceedings of the 37th Meeting of ACL, pages 527–534, College Park, MD, June.

[12] Resnik, P. and Elkiss, A. (2003) The Linguist's Search Engine: Getting Started Guide. Technical Report: LAMP-TR-108/CS-TR-4541/UMIACS-TR-2003-109, University of Maryland, College Park, November 2003.

[13] Say B., Zeyrek D., Oflazer K. & Özge U. (2002). Development of a Corpus and a Treebank for Present-day Written Turkish. Current Research in Turkish Linguistics. Kamile İmer and Gürkan Doğan (eds), pp. 183-192, Proceedings of 11th International Conference on Turkish Linguistics. Eastern Mediterranean University, Northern Cyprus.

[14] Sharoff S. (2006). Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, WaCky! Working papers on the Web as Corpus. Gedit, Bologna