Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 427 – 436

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

An economic approach to big data in a minority language

Šandor Dembitz[a]*, Gordan Gledec[a], Mladen Sokele[b]

[a]University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, HR-10000 Zagreb, Croatia
[b]Croatian Telecom, Planning and Project Management, Savska cesta 32, HR-10000 Zagreb, Croatia

Abstract

Google's n-gram project brought recently big data benefits to several main world languages, like English, Chinese etc. Any attempt to derive such systems, aimed to accelerate the development of NLP applications for world minority languages, in the manner in which it has been done in the project, encounters many obstacles. This paper presents an innovative and economic approach to large-scale n-gram system creation applied to the Croatian language case. Instead of using the Web as the world's biggest text repository, our process of n-gram collection relies on the Croatian academic online spellchecker *Hascheck*, a language service publicly available since 1993 and popular worldwide. The service has already processed a corpus whose size exceeds the size of the Croatian web-corpus created in recent years. Contrary to the Google n-gram systems, where cutoff criteria were applied, our n-gram filtering is based on dictionary criteria. This resulted in a system comparable in size to the largest n-gram systems of today. Because of the reliance on a service in constant use, the Croatian n-gram system is a dynamic one, unique among the systems compared. The importance of having an n-gram infrastructure for rapid breakthroughs in new application areas is also exemplified in the paper.

Peer-review under responsibility of KES International.

Keywords: Croatian language; language modeling; lexical n-gram; n-gram count comparison; traffic modeling.

* Corresponding author. Tel.: +385-1-6129-760; fax: +385-1-6129-616.
E-mail address: sandor.dembitz@fer.hr.

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.123

1. Introduction

The big data trend in the area of natural language processing (NLP) is well expressed in concluding remarks of the Google research team (p. 12 in[1]), which can be summarized in six words: *More words and less linguistic annotation*! However, publicly available large-scale n-gram systems are still the privilege of only 11 Indo-European languages[2,3], the Chinese[4] and the Japanese language[5]. In all cases the WaC (Web as Corpus) approach to big data collection was applied. The WaC trend was followed by South Slavic computational linguists too, which have created recently the corpora for Croatian and Slovene language[6]. In this specific case one must allow for the closeness of the South Slavic languages. The amount of texts written in neighboring languages (especially close to each other are those derived from the former Serbo-Croatian language) within a preselected set of HTML documents is not negligible, and there is no simple and effective way to filter them out, in order to create a "clean" web-corpus for a desired South Slavic language (the standard language identification procedure based on word filters does not help). As far as we know, the Croatian WaC is still in a cleaning process, three years after its creation.

Aware of the obstacles in the case of South Slavic WaC approach, we took advantage of the already operating Croatian academic online spellchecker *Hascheck*[7] and started collecting Croatian n-grams, n = 1, 2,…, 5, a year after the appearance of Brants' and Franz's English n-gram system[2]. The original intention was to use n-grams as the basis for upgrading *Hascheck* into a contextual spellchecker, but in the course of development it became clear that the results are much more broadly applicable. From a respectable amount of data collected so far, we succeeded in developing a consistent, maintainable and upgradable n-gram system, comparable in size to the largest Google n-gram systems.

The paper is organized into five sections. Section 2 gives an overview of *Hascheck* as a tool for big data collection, and some characteristics of its traffic. The n-gram system creation and maintenance, with an insight into system's main properties, and its comparison with the biggest Google systems, is described in Section 3. Section 4 is devoted to rapid breakthroughs in several new NLP application areas, which became feasible because of the n-gram system existence. Finally, Section 5 contains our concluding remarks.

2. Croatian big data tool

Hascheck started as an e-mail embedded service in 1993, at first only for the staff of the Faculty of Electrical Engineering and Computing in Zagreb, but in March 1994 it became a public service, primarily dedicated to the Croatian academic community. In the summer of 2003 the e-mail service was converted into a web service available at http://hascheck.tel.fer.hr/. With the web interface, *Hascheck* became a service adopted worldwide. In the last ten year of operation, the service was used by more than 450 000 users (HTTP cookies) from 124 IP-domains (121 country and 3 generic top-level domains). They have sent more than 7.2 million texts to proofreading, which formed a corpus of more than 1.8 billion tokens (Gtokens).

Hascheck has two subsystems: the real-time subsystem, which spellchecks received text, and the post-processing subsystem, which uses collected process data and performs learning, system statistics and similar tasks. The outcome of learning is the update of the dictionary, i.e. the improvement of the spellchecker's functionality. The learning system incorporated into the post-processing subsystem is what makes *Hascheck* different from other spellcheckers[7].

Hascheck's dictionary is organized into three word-list files:

- a Word-Type (WT) file,
- a Name-Type (NT) file,
- an English-Type (EngT) file.

The WT-file contains Croatian common word-types, words which may occur written in lower-case only, with an initial upper-case letter (at the start of a sentence, for example), or written in upper-case only, and which were not borrowed (with their orthography) from foreign languages, but belong intrinsically to the Croatian language itself. The WT-file started with approximately 100 000 entries, but due to the learning it has increased to more than million word-types. Croatian is an inflected language able to produce, from a changeable word, many more word-forms than