Effective Corpus Virtualization

Miloš Jakubíček[‡†], Adam Kilgarriff[†], Pavel Rychlý[‡†]
‡ NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic
† Lexical Computing Ltd., Brighton, United Kingdom
jak@fi.muni.cz,adam@lexmasterclass.com,pary@fi.muni.cz

Abstract

In this paper we describe an implementation of corpus virtualization within the Manatee corpus management system. Under *corpus virtualization* we understand logical manipulation with corpora or their parts grouping them into new (virtual) corpora. We discuss the motivation for such a setup in detail and show space and time efficiency of this approach evaluated on a 11 billion word corpus of Spanish.

Keywords: corpus, corpus linguistics, virtualization, indexing, database

1. Introduction

This paper brings together two notions from widely separated areas: *virtualization* and *corpora*. Text corpora – large collections of electronic texts – are one of the essential resources for linguistics and have a firm place within computational linguistics and natural language processing. They have applications in a wide range of fields, providing reliable evidence for linguists and statistical models for engineers.

Virtualization has become one of the technology buzzwords of the beginning millenium as more and more people were seeking for appropriate solution for managing large scale IT resources. They were in place and ready to be used but often the inability to carry out reliable predictions that would help distributing the resources among the related services turned out to be a big problem. Using a resource is always a commitment, within the IT field much related to money investments and so questions like: "Does this server need 2, 4, 8 gigs of memory and the other one less, or vice versa, or should we just buy more?" gained a lot of importance, since the right answer led to large savings.

Since better predictions were not really available on the dynamic IT market, people aimed at a technological solution that would allow them to postpone their commitments or not to do them at all; and so we soon witnessed the take up of virtualization starting with processor and memory virtualization allowing a single physical system to host a number of virtual ones and distribute resources among them, continuning with storage virtualization and finally creating a sole market of cloud services.

Current situation in corpus linguistics is to some extent similar to that in IT before virtualization: for many languages there are large text collections available (see e.g. (Jakubíček et al., 2013a; Callan et al., 2009; Pomikálek et al., 2012)) and one has to decide how these will be organized into corpora at the technical level, i.e. as independent database units resulting from a (possibly costly, both in terms of runtime and final space occupation) indexing procedure.

While we presume that the corpus axiom is: the bigger the better, clearly having just a single huge corpus per language is not always desirable for obvious practical reasons – smaller data is always faster to process, one corpus implies one annotation scheme which would then be very limited, and finally one might just find himself in a situation where the subject of studies would be a portion of the language (possibly defined using complex constraints).

The obvious solution to this problem lying in creating separate and independent corpora for any combination of needs becomes less and less feasible for very large datasets. Therefore in this paper we would like to introduce the concept of corpus virtualization, a method allowing flexible management of corpora into logical units, as implemented within the Manatee corpus management system (Rychlý, 2007) used in the Sketch Engine (Kilgarriff et al., 2004).

The structure of the paper is as follows: in the next section we briefly describe the Manatee corpus management system and its past approaches to corpus organization, then we present the approach based on virtualization and its evaluation on a sample dataset.

2. Manatee

Manatee (Rychlý, 2000; Rychlý, 2007) is an all-in-one corpus management system specifically designed for text corpora. As any database system its elementary components can be divided into those that are used for compiling (indexing, building) the corpus (database) index files and those that are then used to query the corpus. Manatee uses a sophisticated set of index files based on the well-known inverted-index approach (Knuth, 1997) allowing complex but fast searching even for complex annotations using the Corpus Query Language (CQL, see(Jakubíček et al., 2010)).

Any reasonable indexing of text data starts with providing an efficient string-to-number mapping of the input words (or lemmas, or tags, etc.) as described in (Jakubíček et al., 2013b). The resulting data structure is called a *lexicon* and allows all other indices to operate on numbers, not on strings, and therefore to be smaller and faster to use.

The corpus consists of three elementary entities: *attributes* (such as word, lemma, tag), *structures* (sentences, paragraphs, documents) and *structure attributes* (metadata on structures, such as document ID), where for any attribute the following indices are compiled:

- attribute text (IDs in the order of appearance in the corpus)
- inverted index (list of positions for each ID)
- lexicon (string $\leftrightarrow$ ID mapping)

corpus	number of tokens (billions)	database size (gigabytes)
esAmTenTen11	8.7	217
esEuTenTen11	2.4	35
esTenTen11	11.1	252

Table 1: Overview of the esTenTen corpus and its parts.

Corpus parts are managed by specifying *subcorpora* for a corpus, where a subcorpus is simply defined (both conceptually and at the technical level) as a set of corpus segments (based e.g. on meta-data annotation). A subcorpus cannot be used as a standalone corpus, it is always accessed only from the main corpus. The subcorpus compilation creates just one index file specifying the subcorpus segments and as for query evaluation, the subcorpus serves just as a filter on top of the full corpus indices.

3. Corpus Virtualization

A *virtual corpus* is defined as a set of segments from one or more corpora. A virtual corpus therefore might be used just for a subcorpus as well if the segments originate from a single corpus – but in most cases this will not be the case and a virtual corpus will rather be a *supercorpus* in this respect. It uses the very same configuration setup as any regular corpus in Manatee except that instead of specifying input text to be processed, a definition file for virtual corpus is given in the following simple format:

```
=bnc
0,1000000
10000000,11000000
=susanne
0,$
=brown
0,1000
```

Each line starting with an equal sign specifies a source corpus to be used, otherwise lines are comma separated position pairs denoting segments to be included into the virtual corpus (where a dollar sign means last corpus position).
This definition file would describe a virtual corpus consisting of the first and eleventh million tokens of the BNC corpus and the whole Susanne corpus and the first 1,000 tokens from the Brown corpus.
While the subcorpus can be seen as a very light-weight concept, a virtual corpus is a heavy-weight mechanism and virtual corpora are first-class databases – they can be accessed without any knowledge of where they come from.
Compilation of a virtual corpus consists mainly of providing a new lexicon and mappings to all existing lexicons of the source corpora. Apart of that only the preexisting indices of those corpora are used for query evaluation resulting in large storage savings while having negligible influence on the query evaluation performance.
We demonstrate the advantages of virtual corpora as opposed to the regular ones on the example of the Spanish esTenTen corpus (Kilgarriff and Renau, 2013) consisting of two substantial parts, the esEuTenTen (European Spanish) and esAmTenTen (American Spanish) as given in Table 1. In Table 2 a comparison of its compilation in both variants is provided: as a virtual corpus consisting of two regular corpora and as a single regular corpus. As can be seen, the virtual corpus approach achieves space savings by factor of more than 20 and time saving by factor of more than 10.

	virtual	regular
space occupied	13 GB	252 GB
compilation time	3.4 hrs	30.6 hrs

Table 2: Comparison of the esTenTen being compiled as a virtual and regular corpus.

4. Related Work

Similar approach to the management issues of large text corpora has been taken within the COSMAS project focusing on the German Reference Corpus (DEREKO, (Kupietz et al., 2010)) where the concept of a virtual corpus suits for selecting working texts out of the whole DEREKO corpus, which would correspond to the subcorpus concept within Manatee. To the authors' best knowledge, the presented approach is unique in that the virtualization operates on entirely independent database entities and the virtualization process creates such a database as result too.

5. Conclusions and Further work

In this paper we justified and presented a method for corpus virtualization within the Manatee corpus management system that is efficient in terms of both savings in compilation time and occupied disk space while having very little footprint on the system performance during query evaluation.
Another exploitation of corpus virtualization currently under development is that for effective paralellization of corpus compilation by dividing the data into n-parts that will be compiled separately, then joined into a virtual corpus and this corpus will be finally devirtualized into a regular one.
All the implemented functionality belongs to the open source part of the Manatee corpus management system released as Manatee-open under the GPLv2 software license at `http://nlp.fi.muni.cz/trac/noske`.

6. Acknowledgements

This work has been partly supported by the Ministry of Education of the Czech Republic within the LINDAT-Clarin project LM2010013.

7. References

Callan, J., Hoy, M., Yoo, C., and Zhao, L. (2009). Clueweb09 data set. Online presentation available at: `http://boston.lti.cs.cmu.edu/classes/11-742/S10-TREC/TREC-Nov19-09.pdf`.

Jakubíček, M., Rychlý, P., Kilgarriff, A., and McCarthy, D. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. In *PACLIC 24 Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 741–747, Tokyo.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013a). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013*, pages 125–127, Lancaster.

Jakubíček, M., Šmerk, P., and Rychlý, P. (2013b). Fast construction of a word-number index for large data. In A. Horák, P. R., editor, *RASLAN 2013 Recent Advances in Slavonic Natural Language Processing*, pages 63–67, Brno. Tribun EU.

Kilgarriff, A. and Renau, I. (2013). esTenTen, a Vast Web Corpus of Peninsular and American Spanish . *Procedia - Social and Behavioral Sciences* , 95(0):12 – 19. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013) .

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, Lorient, France. Universite de Bretagne-Sud.

Knuth, D. E. (1997). Retrieval on secondary keys. *The art of computer programming: Sorting and Searching*, 3:550–567.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis and Mike Rosner and Daniel Tapias, editor, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta, Malta, May. European Language Resources Association (ELRA).

Pomikálek, J., Jakubíček, M., and Rychlý, P. (2012). Building a 70 billion word corpus of English from ClueWeb. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis, editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace*. PhD Thesis, Masaryk University, Brno.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno.