Zawilinski: A library for studying grammar in Wiktionary

Zachary Kurmas
School of Computing
Grand Valley State University
kurmasz@gvsu.edu

ABSTRACT

We present Zawilinski, a Java library that supports the extraction and analysis of grammatical data in Wiktionary. Zawilinski can efficiently (1) filter Wiktionary for content pertaining to a specified language, and (2) extract a word's inflections from its Wiktionary entry. We have thus far used Zawilinski to (1) measure the correctness of the inflections for a subset of the Polish words in the English Wiktionary and to (2) show that this grammatical data is very stable. (Only 131 out of 4748 Polish words have had their inflection data corrected.) We also explain Zawilinski's key features and discuss how it can be used to simplify the development of additional grammar-based analyses.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces

Keywords

wiki, Wiktionary, MediaWiki, parse, language, inflection

1. INTRODUCTION

This poster presents the design of *Zawilinski*,[1] a Java library for extracting and analyzing inflection data in Wiktionary (i.e., "word endings"). It also demonstrates Zawilinski's usefulness by presenting the results of our Zawilinski-based analysis of the inflections of Polish words in the English Wiktionary. (The Polish Wiktionary contains less inflection data than the English Wiktionary.)

Although the accuracy of grammatical data, obviously, does not apply directly to wikis overall, the study of objective data sets does provide a perspective that can be combined with other wiki analyses (see Section 4) to produce an increasingly accurate and useful understanding of wiki accuracy and contributor behavior.

Not only can Zawilinski be quickly extended to support the analysis of other languages, but it can also be used to support the analysis of any objective, consistently formatted data in a MediaWiki XML dump.

2. ZAWILINSKI

Zawilinski currently has three key features:

1. It can filter large MediaWiki XML documents and retain only those `<page>`s and/or `<revision>`s that contain data of interest.
2. It can load a MediaWiki XML document and present the contents as a tree of Java objects. Technically speaking, Zawilinski uses JAXB to *unmarshal* the XML document into Java content objects.[2]
3. Once a Wiktionary-based XML document is loaded, Zawilinski can search the `Revision` objects for templates containing the words' inflections, expand those templates, then create Java objects containing all of a word's inflections.

Typical MediaWiki dumps are too large to be completely loaded into memory. Zawilinski uses a novel combination of pre- and post-filters to easily and efficiently remove those elements that do not contain data of interest. As soon as it has unmarshalled a `<page>` or `<revision>` element, a *post-filter* decides whether to retain or discard the resulting Java object. Zawilinski also allows users to place *pre-filters* before the unmarshaller to remove obviously unnecessary text. (Technically speaking, these pre-filters are SAX filters — subclasss of `XMLFilterImpl`.) Pre-filters mitigate the inefficiency of unmarshalling a large element only to discard the resulting object. Also, pre-filters prevent excessively large elements from causing Java OutOfMemory errors.

This combination of pre- and post-filters allows Zawilinski to leverage both the efficiency of filtering data before it is loaded into a Java object and the simplicity of examining an entire MediaWiki `<page>` as a tree of Java objects. (More technically, Zawilinski combines the efficiency of a SAX filter with the simplicity and convenience of examining Java objects after they have been unmarshalled by JAXB.)

Although we use Zawilinski specifically to extract inflection data from Wiktionary, we designed it to be useful for

[1] Roman Zawiliński (1855 – 1932) was a Polish linguist, pedagogue and ethnographer. See `http://en.wikipedia.org/wiki/Roman_Zawili%C5%84ski`.

[2] `http://java.sun.com/developer/technicalArticles/WebServices/jaxb/`

WikiSym'10, July 7-10, 2010, Gdańsk, Poland.
Copyright 2010 ACM 978-1-4503-0056-8/10/07 ...$10.00.

a wide variety of MediaWiki-based research. For example, studying pronunciation or etymological data requires no changes to the loading and filtering classes. Users need only write code to find and extract the desired data from the text of a loaded `Revision` object. To study other objective, consistently formatted data sets in a MediaWiki site, users need only implement the appropriate pre- and post-filters to produce an XML document of manageable size, then write the code to extract the desired data.

3. PRELIMINARY RESULTS

We present a sample of the results of using Zawilinski to analyze the inflection data for Polish words in Wiktionary. Generating the results required only 300 lines of Java source code outside Zawilinski. Although the results below do provide insight into the behavior of wiki contributors, the primary purpose of these results is to demonstrate the ease and speed with which researchers can use Zawilinski.

We analyzed the Wiktionary history dump from 20 February 2010. Of the 7453 Polish words in the English Wiktionary, 5508 contain inflection data: 4373 nouns, 375 verbs, and 760 adjectives. The remaining words either don't inflect (e.g., adverbs, prepositions), or haven't yet had their inflections entered into Wiktionary.

Stability: Of the 4748 nouns and verbs, 208 have had their inflection data updated. However, only 131 of these updates are "true" corrections. The remaining 77 updates add more inflections without modifying existing data. Of the 131 "true" corrections, only 10 were later corrected again.

Correctness: Given an electronically accessible authoritative source of grammatical data, Zawilinski can compute the correctness of the inflections in Wiktionary. No such resources yet exist for Polish; however, there are two manually accessible authoritative sources:[3] *Słownik gramatyczny języka polskiego* (SGJP: The Grammatical Dictionary of the Polish Language) is an electronic dictionary containing the inflections of over 245,000 Polish words [4]. In addition, the University of Pittsburgh maintains an online dictionary that also contains inflection tables for many Polish words.[4]

To demonstrate Zawilinski's ability to measure the correctness of inflection data, we compared two small sets of Wiktionary data to our authoritative sources: The 375 Polish verbs that currently have Wiktionary entries, and the 131 words that have been corrected.

Verb correctness: Of the 375 verb entries, only 19 have conjugation data that differ from our authoritative sources.

Correctness of updates: Of the 131 words that were corrected, only 11 of the words still contain errors in the current Wiktionary. Of the 150 corrections (some words were corrected more than once), only 10 updates replaced correct data with incorrect data. All but one of these incorrect updates were later corrected.

4. RELATED WORK

Related tools: Zawilinski is one of several tools that parse MediaWiki XML dumps. Other popular tools include JWPL[5], JWTKL, mwdumper,[6] and wikixmlj. [7] Space constraints prevent us from discussing each tool in detail; however, Zawilinski differs from these (and other) tools in two key areas: (1) It does not require the XML document to be initially loaded into mySQL or some other database. (2) It can save the filtered XML document as a smaller XML document for later use.

Related Analyses: In 2005, the journal *Nature* found that, among randomly chosen science articles, the Wikipedia entries averaged only one more mistake than the corresponding Encyclopedia Britannica entry [2]. Other studies have attempted to ascertain an article's accuracy based on secondary factors, such as the author's reputation. Adler and de Alfaro evaluated an author's reputation based on how quickly his or her edits were updated or rolled back [1]. Similarly Hu et al. estimate an author's authority based on the number of words that survive future edits [3]. Zawilinski complements this work because (given objective data) it allows us to correlate the secondary measures presented in [1] and [3] with the actual correctness of the data. Thus, we can measure the degree to which longevity of an author's updates correlates with the accuracy of his or her data.

5. CONCLUSION

Zawilinski is a powerful and flexible library to support the analysis of objective data in Wiktionary and other MediaWiki sites. We have presented its key features and demonstrated that it can be used to quickly analyze a language's inflection data in Wiktionary. We look forward to performing a comprehensive analysis on all Polish words in Wiktionary, analyzing other languages as authoritative sources become available, and repeating some of the aforementioned related studies and correlating the results with correctness.

`Zawilinski` is available on the web at `http://www.cis.gvsu.edu/~kurmasz/Zawilinski`.

Acknowledgments

We would like to thank the many faculty and students who contributed to this project including: Robert Adams, Jerry Scripps, Christine Rydel, Eva Lewak, Ira Woodring, Ryan Lammers, Sean Fisk, Chris Sechler, and Nick Butcher.

6. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system. for the wikipedia. In *Proceedings of the World Wide Web Conference*, May 2007.

[2] J. Giles. Internet encyclopaedias go head to head. *Naure*, 438:900–901, December 2005.

[3] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: Models and evaluation. In *Proceedings of the Conference on Information and Knowledge Management*, 2006.

[4] Z. Saloni, W. Gruszczyński, M. Woliński, and R. Wołosz. *Słownik gramatyczny jęnzyka polskiego.* Wiedza Powszechna, 2007.

[3] The licenses of these sources prohibit users from automatically extracting data (e.g., through the use of a computer program). The expectation is that users make individual queries by hand.

[4] See `http://polish.slavic.pitt.edu/~swan/beta/`.

[5] `http://www.ukp.tu-darmstadt.de/software/jwpl/`

[6] `http://www.mediawiki.org/wiki/Mwdumper`

[7] `http://code.google.com/p/wikixmlj/`