Storing morphology information in a wiki[1]

Radovan Garabík
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
813 64 Bratislava, Slovakia
korpus@korpus.juls.savba.sk, http://korpus.juls.savba.sk

Abstract. There are different ways of storing information about morphology. We describe the way of organizing morphology data in a form suitable to be kept as plain text files inside of a MoinMoin wiki engine and the practical results of keeping information about Slovak morphology.

Keywords: morphology analyser, wiki, MoinMoin, tagset, part of speech, scalability

Introduction

We successfully developed the Slovak language morphology analyser based on Levenshtein edit operations [1, 2]. The original aim was to cover all the words present in the Short Dictionary of the Slovak Language (KSSJ) [3] and some additional frequent words. The Levenshtein edit operation based paradigm classes proved very useful for quick semi-automatized construction of all the word forms derived from a given lemma. However, as the number of words included reached the originally projected goal, the project entered a maintenance phase with new words being added only sporadically with focus towards long term storage and reutilization of the data, and consequently a new approach appeared to be desirable.

Wiki software

We settled on the MoinMoin[4] wiki engine (presently we are using version 1.6.3). MoinMoin is a wiki written completely in the Python programming language [5] using flat text files as a storage backend rather than a database. This makes it particularly attractive for our needs because of the programming language involved and the ease of making various data modifications and extraction using just common text processing tools. MoinMoin is also fully Unicode aware, and all the stored data and I/O are invariably in UTF-8 encoding.

MoinMoin contains a built-in full text search engine or it can use the Xapian libraries[6].

[1] This work was supported under the project MONDILEX – Conceptual Modelling of Networking of Centres for High-Quality Research in Slavic Lexicography and Their Digital Resources

MoinMoin can be extended by writing macros or plugins – in particular, we extended the default parser to display the morphology tags in better human-readable forms (with explanation of different grammar categories), while keeping the original data intact and in a terse form suitable for computerized parsing.

MoinMoin also supports XML-RPC access to the data, a feature that can be potentially interesting in view of eventual integration of the database into external linguistic resources.

Our MoinMoin server runs on a modest Intel Pentium 4 server with 2 GB of RAM, two IDE disks assembled into a RAID 1. The operating system is GNU/Linux (Ubuntu Hoary) and the wiki data is kept on a ReiserFS filesystem.

Data structure

Although the primary purpose of the wiki is to keep the data for the automatized NLP processing purposes, we still aim for the data to be useful also as a reference database for dictionary-like queries, and therefore the design of the pages is made with this goal in mind.

The basic unit of the wiki data is called a page (using MoinMoin terminology). Each page contains data pertaining to one lexeme, i.e. a lemma with full paradigm and morphology annotation. Each page name is equal to the lemma taking into account common capitalization of words in Slovak (proper nouns)[2]. In the case of lexical homonymy, pages are named by the lemmas with the part of speech tag attached in parentheses[3].

We strived to keep the page structure to be both human-readable and human-editable as well as being easy to parse automatically.

The page body has a form like this:

[2] An important point, because by design the final morphology analyser disregards the capital letters and gives all the lemmas in lowercase.

[3] E.g. *mať_(V)* for a verb, *mať_(S)* for a noun.

```
== Lema ==
ucho

== Paradigma ==
SSns1: ucho
SSns2: ucha
SSns3: uchu
SSns4: ucho
SSns5: ucho
SSns6: uchu
SSns7: uchom
SSnp1: uši, uchá
SSnp2: úch, ušú, uší
SSnp3: ušiam, uchám
SSnp4: uši, uchá
SSnp5: uši, uchá
SSnp6: uchách, ušiach
SSnp7: ušami, uchami
----
[[Kategória:Substantíva]]
```

Text 1: Example of a wiki page (lemma *ucho*).

The page body contains several sections. The first one is the *Lema*, which contains just one word, the lemma. Then the *Paradigma* section follows, containing the inflectional paradigm spelt out in full. For each grammar category there is one corresponding line, with the morphological tag separated from the form by a colon (:). Alternative forms per one grammar category can be either given on a separate line, or on the same line, separated by a comma (,). At the end of a page there is the part of speech category the described word belongs to.

Homonymy

We are talking here about the basic homonymy only, where lemmas for two different words (two different parts of speech) are identical. The other forms of homonymy (inflectional) are automatically taken care of by keeping the homonyms under their corresponding lemmas and morphology tags.

In case of part of speech homonymy, we create a special disambiguation page linking to all the possible lemmas.

```
== Lema ==
mať

== Pozri ==
[[mať_(S)]] [[mať_(V)]]
----
[[Kategória:Dezambiguácia]]
```

Text 2: Example of a disambiguation wiki page (lemma *mať*).

Reflexive verbs

In Slovak, reflexive verbs [7] are marked by a special separate morpheme *sa/si*, which is separated from the verb and has relative freedom of movement around the verb[4]. As there exists a reflexive/non-reflexive dichotomy (i.e. reflexive verbs almost always have their non reflexive counterpart), we decided to keep only the non reflexive parts in the dictionary, without the *sa/si* pronoun. Several singular cases of reflexive verbs without a meaningful standalone non-reflexive counterpart (*smiať sa, báť sa, uvedomiť si, čudovať sa*) do not pose any problem – the missing *sa* is confusing only for the uninitiated users.

Traditionally, *sa* and *si* are called “reflexive pronouns” if semantically there is a discernible action performed on the agent (i.e. they can be seen as contractions of personal pronouns *seba* and *sebe*), otherwise they are considered to be a part of a verb. This is just a convention – we could denote them equally good as particles, indeed this is how they are sometimes classified in the traditional Czech grammars. We simplified our task by assigning the *sa* and *si* a special morphology tag **R**, regardless of their semantic use.

Part of speech distribution

Currently, the wiki contains 77567 entries. Categorised by the POS type, we have the following distributions:

[4] Unlike other languages, e.g. in Russian the reflexive pronoun/particle takes a form of a clitic inseparably bound to the verb.

28163	verbs
26061	substantives
13100	adjectives
5069	adverbs
1297	abbreviations
1104	participles
656	interjections
369	particles
369	pronouns
311	numerals
123	prepositions
110	conjunctions
72	citation elements[5]
26	part of multiword expression[6]
2	*sa/si*
1	*by*[7]
716	disambiguation pages

Table 1: Distribution of parts of speech

Scalability

As the total amount of entries in the database reaches tens of thousands, with the possibility of growth up to several times the number, it is important to achieve reasonable scalability of the wiki engine. Since the MoinMoin stores each page in its own directory and all the directories are stored under one parent directory, it is important for the underlaying file system to be able to cope with many thousands of entries per directory. All the major modern Linux file systems [8] have no problems with this usage pattern. Probably the best filesystem for these purposes at the moment is ReiserFS, which also has other convenient features such as tail-packing to conserve disk space, since the files used by the backend storage are predominantly way below file system block size. The total size of our data is 1.2 GB of disk storage.

[5] “Citation element” is a foreign language word appearing in Slovak text, e.g. most often in book or movie names, or French or Latin quotations. In our wiki, only a few such words are included.

[6] Used to mark standalone morphemes that are a part of multiword expressions – these are in fact just a remnant of our tokenization.

[7] Special conditional morpheme, traditionally classified as a particle.

Basic usage works well, and direct searching for a lemma, page editing, revision history and similar actions are performed without noticeable delays. However, the built-in full text search engine is unable to cope with the amount of data. Basic searches for an inflected word form typically take many long minutes of 100 % CPU utilization. After the switch to the Xapian search engine, the search for a word form is instantaneous. However, other features that depend on the number of pages are difficult to use, e.g. displaying all the pages in one category takes several minutes (much of the time is not due to searching, but to formatting such a huge number of links).

Usage

The wiki can be used directly as a reference dictionary of inflectional data. However, we are using it mostly as a source of data for a morphology analyser, transforming the data from the wiki into constant database tables [9] for quick retrieval, further independent of the wiki software.

We also convert the data into a nicer looking format for the DICT server (RFC 2229) [10] for a quick web-based search, integrated with several other Slovak language dictionaries.

Conclusion

Storing rich morphology information on the level of tens of thousands of words into a MoinMoin wiki based system is viable, as long as special care is taken not to use features that scale badly with increasing number of pages such as Category pages – in our wiki containing just a static description of each part of the speech category, not the list of all pages belonging to a given category. The wiki is used as a source of data for various morphology related automatized tasks, as well as a source for a human-readable dictionary of Slovak morphology.

References

1. Garabík, R. (2006). Slovak morphology analyzer based on Levenshtein edit operations. In: Proceedings of the WIKT'06 conference, pp. 2—5. Bratislava, Slovakia.
2. Левенштейн, В. И. (1965). Двоичные коды с исправлением выпадений, вставок и замещений символов, Докл. АН СССР, 163, 4, pp. 845—848.
3. Krátky slovník slovenského jazyka. Ed. M. Považaj. Veda, Bratislava, 2003.
4. http://moinmo.in
5. http://www.python.org

6. http://www.xapian.org
7. Dvonč, L. et al. (1966). Morfológia slovenského jazyka. Vydavateľstvo Slovenskej akadémie vied, Bratislava. pp. 377—388.
8. Piszcz, J. (2006). Benchmarking Filesystems Part II. Linux Gazette, 122.
9. http://cr.yp.to/cdb.html
10. Faith, R., Martin, B. (1997). “A Dictionary Server Protocol”, Request for Comments 2229, Network Working Group, ftp://ftp.isi.edu/in-notes/rfc2229.txt