English-Russian Machine Translation
A Grammatical Approach

by

Edward Charles Eberle

October 9, 2017

Edward Charles Eberle
October 9, 2017

English-Russian Machine Translation
San Diego, California USA
A Grammatical Approach

Table of Contents

Edward Charles Eberle
October 9, 2017

English-Russian Machine Translation
San Diego, California USA
A Grammatical Approach

Introduction

I have written a prototype for an English-Russian machine translation program and I thought some explanation necessary, hence this paper. My biggest complaint with machine translations is not the accuracy of the content, it is the a-grammatical presentation of the output, and I hope, at the very least, to provide an approach that will solve this problem on the English end. The approach I have used is based on my 30 years of foreign language study and the seven languages I have literacy in. I have a Master of Arts in Russian Literature from San Diego State University (1997) and am current in Russian, French, German, Spanish, Polish, Vietnamese and of course English. I have spent the last 20 years studying computer science on my own and have decided to use Python 3.x as the language for the application, PyQt5 for the graphical user interface and PostgreSQL for the database. The code is written as open source code and will be available for use and emendation by all and sundry. On to the the approach.

I initially created an English language categorization scheme taken largely from the website www.englishclub.com and various others I didn't document. It is contained in the file “wordtype.py” in the machinetrans/data directory and goes something like this:

Basic Word Groupings

noun, adjective, verb, adverb, interjection, pronoun, article, preposition, symbol, conjunction, interrogative, participle, invariant

The reason the participles have their own category is because the Russian language has a rather extensive collection of participial constructs, something the writer Nabokov called “adverbish mongrels,” and I hope to capture some of this high literary content with my program.

If you really want to see Russian participles in all their glory read Leo Tolstoy's *War and Peace* in the original Russian. Invariant corresponds to the Russian category of the same name, which I am using for the subjuntive “бы” , acronyms, any foreign terms and an “other” category which I hope will not be too much abused. I have further broken these categories into something I have called “varieties” as follows:

Nouns

abstract, proper, concrete, collective, compound

Verbs

motion, other

Adjectives

descriptive, comparative, superlative, short adjective

Adverbs

state, manner, place, time, degree, other

Pronouns

personal, demonstrative, possessive, interrogative, reflexive, reciprocal, indefinite, relative, inclusive, variety

Participles

present active, present passive, past active, past passive, present verbal adverb, past verbal adverb 1, past verbal adverb 2, short past passive, formal imperative, informal imperative

Prepositions

place, time, other

Conjunctions

coordinating, subordinating

Symbols

glyph, number, other

Interrogatives

person, place, time, method, reason, ownership

Invariants

subjunctive, foreign, acronym, other

These are my word “varieties,” needless to say this is mostly speculation at this point and subject to revision. You will note that there are both interrogative pronouns and interrogatives as a category, this is a Russian feature and technically English interrogatives are Russian adverbs. These types are implemented as Python “tuples” with the data type string and can be quite easily used as strings by the program. Furthermore, “tuples” are ordered, and can be sorted using the “sorted” method for the menus so the user does not have to hunt each time for the given value needed. Note that as far as I know the categories article and interjection are almost vestigial, while interjections have no variety at all. In the category symbol, glyph is a non-English character and number speaks for itself.

There are subcategories present, and there will probably be more as time goes on.

Person

first, second, third, plural first, plural second, plural third

Tense

imperative, simple present, simple past, perfect simple past, perfect continuous past, continuous past, continuous present, perfect simple present, perfect continous present, simple future, continuous future, perfect simple future, perfect continuous future

Person presents difficulties because Russian, following its Latinate brethren, has an informal you and a formal you, and the formal you can be plural, thus the “plural second” in this category, this is almost always lost in translation. You will note that there are twelve tenses. This huge amount will hopefully cover both the English situation and the Russian adequately. In Russian we meet a challenge in the verbs of motion, there are verbs for general motion in the present and past tense, directed motion in the present and past tense and a plethora of various kinds of motion being distributed among walking, driving, carrying, etc. These show up in translation as “to drive away”, to carry away”, “to walk past”, “to fly past”, each one being a single Russian verb. Add to this that there are continuous future tenses provided by combining verbs with the future forms of “to be.” I haven't quite reached this traffic snarl but it looms in the future. Among the other verbs there is a distinction between ongoing and completed action, ongoing occurring in the present and past and completed occurring in the future and past, I have defined a two element category for verbs consisting of:

imperfective

imperfective, perfective

Other two element categories are:

Noun Type

location, thing

Animate

animate, inanimate

The noun you cannot lift is obviously a location. And the book that doesn't walk off is inanimate.

Then there are the rest:

Gender

masculine, feminine, nueter, plural

Case

nominative, accusative, genitive, dative, instrumental, prepositional

Object Case

nominative, accusative, genitive, dative, instrumental, prepositional, none

Which mirror the ancient Latin declensions and apply to nouns, adjectives, pronouns and participles in the Russian.

I hope to keep this file as small as possible. The advantage of having all this in one file as a static data type is huge. You can make menus for data entry. You can quickly add to an SQL command the details of your current word, and you can limit word searches to existing categories. Also making program wide changes is much easier.

Did I say SQL? I have implemented my data as a PostgreSQL database. The advantages are that if you have any need to find a given word using two or three entries from columns in the table it occupies to define what it is, it is not a problem. Also updating data is more economical because all you have to do is ask for "DISTINCT" data and all the duplications are removed. You will find the template for the database in the file "machinetrans/dataentry/langdatacreate.py."

Grammar and English Production

What I am doing currently is generating English sentences from semi-random data. This is my bid to create grammatical English sentences from the Russian language which has no set word order. I am using sentence diagrams to create sentence templates for generation of English sentences (see the files with ...gensql in their name). The ...gensql files are subclasses of the "wordgensql.py" file that digs up all the Enlglish terms and puts them in lists to be doled out randomly to the various construct files have names like subject.py and participle.py by the ...gensql files. The ...gensql files are designed as some basic classes that handle subject, objects, and participles. For example the subject class takes an obligatory subject and verb and allows for optional article, adjective and/or adverb. It is my hope that by using this approach I will be able to look at the final product without pulling out my red pen to correct bad grammar (I did a little teaching too). By semi-random data I mean that using the categories described above and a random number generator to specify which one of a given category is chosen, the data is provide by the sentence classes. If you really want to see the code, look in the file “wordgensql.py” as it is the base class for all my sentence generation.

Finally there exists a file that deserves a fanfare for its introduction, it is “machinetrans/data/wordmorph.py” This class encapsulates all the declensions and conjugations of the Russian language. I use this file to automate the entry of Russian wordforms, a single verb can contain up to 105 variations, a noun has at least 12 forms and an adjective 24. Entering data in all its combinations and permutations is a bit easier with this data encoded. I have tried to designed the program so that adding a new declension or conjugation only requires a small adjustment in the classes that use it. I hope to one day make the program declensions and conjugations, except perhaps for participles, extensible so you can just add to the list and go.

I am indebted to the folks at www.python.org (Python) and www.qt.io (Qt), as well as www.riverbankcomputing.com (PyQt) and, of course, www.postgresql.org (PostgreSQL). All of these sites provided me with the open source software needed to create the programs involved, and the documentation for each package that is essential, and don't forget www.englishclub.com and an ever present ghost in all this www.wiktionary.org, which is an on-line multilingual dictionary.

Bibliography

www.python.org. Source of the Python language used to create the program.

www.qt.io. Source of the Qt windowing system.

www.riverbankcomputing.com. Source of the Python version of Qt.

www.postgresql.org. Source of the PostgreSQL database used in the program.

www.englishclub.com. Source of the language categorization scheme.

www.wiktionary.org. Online dictionary that is often used in developing the vocabulary.

Ben T. Clark. *Russian for Americans.* New York: Harper & Row Publlishers, 1973.

Ed. Paul Falla, Marcus Wheeler, Boris Unbegaun, Colin Howlett. *The Oxford Russian-English Dictionary*. Oxford: Oxford University Press, 1993.

Pulkina, I. M. *A Short Russian Reference Grammar: With a Chapter on Pronunciation*. Ed. Dr. P. S. Kuznetsov. Moscow: Foreign Language Publishing House, 1966.