SLAVONIC NAMED ENTITIES IN GATE

Elena Paskaleva, Galia Angelova, Milena Yankova, Kalina Bontcheva, Hamish Cunningham, Yorick Wilks
This work is achieved within the project BIS-21 BULGARIAN INFORMATION SOCIETY, CENTER OF EXCELLENCE FOR EDUCATION, SCIENCE AND TECHNOLOGY IN 21 CENTURY (ICA1-2000-70016), where the Department of Linguistic Modeling of the Central Laboratory for Parallel Processing, Bulgarian Academy of Sciences has the Computer Science Department of Sheffield University as a twinning partner in Work Package WP4: Human Language Technologies for Slavonic Languages, part 1 - "Information Extraction (IE) for Slavonic languages”. The aim of the WP 4.1 is to transfer the Sheffield's leading expertise in IE to Slavonic languages; adapt for Bulgarian and further elaborate if necessary the existing GATE IE modules for English.

The first stage of this elaboration and tuning to the specificity of Slavonic languages of the GATE IE modules is related to procedures of Named Entities extraction.

1. Introduction

One of the Information Extraction sub-tasks is the recognition of the Named Entities (NE) in the raw texts. Named Entity recognition involves processing a text and identifying certain occurrences of words and expressions as belonging to particular categories of Named Entity (NE) [Cun99c]. These categories are locations, persons, organizations, dates, times, monetary amounts and percentages. This information can be used to tag a text for categorisation, but can also be used to support automatic text summarisation, information retrieval, etc.

The Bulgarian NE modules discussed here employ a conventional method that decomposes the input text into words and extracts each Named Entity by referencing gazetteer lists and applying pattern-matching rules.

We have built the NE recogniser for Bulgarian using three main processing resources: a tokeniser, a gazetteer and a finite state transduction grammar written using JAPE [Cun02a]. These modules are built within version 2 of Sheffield’s language engineering framework GATE - General Architecture for Text Engineering [Cun02b,Cun02c]. The modules communicate via GATE’s annotation API, which is a
directed graph of arcs bearing arbitrary feature/value data, and nodes rooting this data into document content (in this case text).

The **tokeniser** splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an capitalised token, all upper case, etc.)

The **gazetteer** consists of lists such as cities, organisations, days of the week, etc. The gazetteer lists are compiled into finite state machines, which can match text tokens.

The **grammars** consists of hand-crafted rules describing patterns to match and annotations to be created as a result. The pattern-action rules are written in JAPE (Java Annotations Pattern Engine) language [Cun02a]. JAPE provides finite state transduction over annotations based on regular expressions. A JAPE grammar consists of a set of phases, each of which consists of a set of pattern/action rules, and which run sequentially. Patterns are specified by describing a specific text string, or annotations previously attached to tokens (e.g. annotations created by the tokeniser or gazetteer).

In this report we discuss the necessary changes the gazetteer’s lists and in the grammar rules for the needs of the Bulgarian module of GATE are discussed.

2. Named entities as linguistic and information units – their general features and distribution in texts

Named Entities (NE) are one of the basic IE backbones and their successful identification gives us a substantial part of the Information that we wish to Extract. The main obstacle for successful identification of the named entities is their nature – an open set, dynamically formed and continuously updated. This specificity of NE makes impossible their preliminary definition in the lexical data base of the system.

The main groups of NE on which the team efforts were concentrated were: **Person Names, Names of Organizations and Dates**.

For the dynamic identification of NE several principles were used, separately or in combination, with a different degree of linguistic support, by using linguistic knowledge from different language levels. In general, the portion of linguistic knowledge is quite poor in these procedures (due to the linguistic nature of NE). The correlation of these
principles and their implementation in the English NE modules in GATE as well in the Bulgarian module version varies for each group of NE and in each concrete language.

2.1 Person names

Person names have high information weight as they denote the participants in the events. Their structure and punctuation particularities strongly depend on the language. Obviously, the first base for the investigation of their nature are the local names. However, this does not mean that the foreign names are of secondary importance. On the contrary, for certain topics the latter are of general importance. In that case their analysis as linguistic unit can be approached from the perspective of using rules of name transcriptions into cyrillic, combined with already existing English NE modules from GATE. For every international news and comments the list of foreign names will be the same in content but written in cyrillic.

A main source for the investigation of Bulgarian person names are the lists of names compiled for other needs. A traditional source of this type is a telephone book. In our experiment it was used as the main repository of linguistic data.

Other significant source of investigation is the text base itself.

On this base the different principles of identification of person names were defined – related to the different levels of description and different portion of knowledge.

Information at alphabetic level

The first and the most surface level of this knowledge which is more printing than linguistic, is the information about letters case and punctuation signs.

The identification of every person name (capitalized as a rule) is very often correlated to the identification of sentence boundary. More precisely, the already set sentence boundary, marked by some strictly defined punctuation signs (most of them ambiguous unfortunately), hampers the identification of the next string in the linear printing sequence as a proper name or as capitalized word (first in the next sentence).
That candidate-NE can be defined as a real-NE through
♦ a pattern matching procedure on the list of person names
♦ its previous occurrences as nonambiguous person name (following a
  non-capitalized string).


On these identification levels the rules are language-independent and can
be applied directly to every text base, including Slavonic one, where
special printing signs denote the sentence boundary and the person name.

Pattern matching in lists

The orientation to a given language and the tuning to its specificity start
with the lists - a pattern matching tool. They are of two basic kinds:
♦ the elements accompanying the names. A core list of them contains
  the specific units preceding the person name – titles, professions and
  others. (Prof, Dr., Mr. etc.)
♦ The person names.

The second type of lists - a basic matching instrument, extended and
updated continuously, is obviously larger.

General principles of recognising person names in the English
NE module

The gazetteer lists for English in GATE deal with two types of proper
names – first and second. The first names, which are not so numerous
and are members of a not so fast extended set, presuppose the
identification of the following capitalized string as a second name. The
second names are not listed, but they are calculated as such, based on the
presence of a first name or a name identifier (e.g., title or profession).
There are 10 200 first names included in the English gazetteer list. A list
of second names is not used as they are recognised using JAPE grammar
rules.

Lists of person names compiled for the needs of Bulgarian NE
Recognition

Before defining the lists of Bulgarian person names – first and family
names and the ways of their treatment for the identification of NE, we
investigated a ready-made list of person names. The telephone book of
Sofia consisting of 330 000 records – combinations of first and family name.

The experiments on that list showed the following correlations between personal and family names:
From the 330 000 names in the list through alphabetisation and sorting were compiled the following sublists:

A. List of family names – 27 500. They are not included in the GATE-BG system lists for matching and serve only as data support for the elaboration of the rule-based identification of the Bulgarian family names. The substitution “list ---> rules” was possible due to the following specificity of Bulgarian name formation:

The family name in Slavonic languages has morphological structure, directly mirroring the semantics of its formation as follows:

\textit{Personal name} + \textit{possessive word derivation suffix} + \textit{gender flexion}

So, the family name \textit{Иванов} has the reading - \textit{the son} of \textit{Иван}, and \textit{Иванова} resp. \textit{The daughter (or wife) of Иван}. See the rule below.

Rule: LastNameOv

// Иванов
// Иванова

```java
{
    \{Token.orth == upperInitial \}
    :person
    -->
    {
        gate.AnnotationSet person =
        (gate.AnnotationSet)bindings.get("person");
        gate.Annotation personAnn =
        (gate.Annotation)person.iterator().next();
        String word = (String)personAnn.getFeatures().get("string");
        gate.FeatureMap features = Factory.newFeatureMap();
        if (word.endsWith("ова") ||
            word.endsWith("ева") ) {
            features.put("gender", "female");
            features.put("rule", "LastNameOv");
            annotations.add(person.firstNode(), person.lastNode(),
            "LastName", features);
        } //if
        else if ( word.endsWith("ов") ||
            word.endsWith("ев") ) {
            features.put("gender", "male");
            features.put("rule", "LastNameOv");
        }
    }
}
```
This distinct structure of the Bulgarian family name allowed us after the alphabetization, sorting and manual filtering of the family names in the telephone book to state that: 91% of the family names can be directly calculated through their morphological components. In this calculation both the family name and its gender are recognised.

The morphological tools for the formation of Bulgarian family names are 6 suffixes. The most frequent is ОВ – 46.4%, followed by ЕВ – 24.4%. The others are: СКИ/a, ИН/a, ШКИ/a, ЧКИ/a. Only 9% of the family names don’t use these flexions but their investigation shows a specific formation for other Slavonic languages as well as Armenian. A deeper investigation on a bigger records base would demonstrate some geographical dependency of the names - i.e. the south Bulgarian sites will contain in their telephone books much more Turkish names. It is to mention here that the Turkish names have also a very strict morphological shape.

In case when morphological rules can identify the family names, the list for matching names can be reduced only to the foreign names.

B. List of first names. They are less than the family ones but are not calculable by morphological rules for the following reasons:
♦ Their morphological structure is not so distinct as the first name is the primary element in the name formation
♦ There is an ambiguity between the personal names and the common words (like Eng. Sunny)

The investigation of the telephone book shows that its 330 000 records contain 6500 unique first names - 3800 female and 2700 male. Within these two subsets 250 female and 230 male ambiguous names were filtered.

In other words: the troubles with the personal names’ identification are related to their formal incalculability and the ambiguity with other linguistic units – the common names.

This fuzzy structure of the personal name imposed the use of list for pattern matching. Its volume is however 4.3 times smaller than the concatenated family names. Another advantage of this list is its relative “closeness” in comparison with the family names. The language is more productive in the formation of the family names than of the personal ones.
It is to note also that the ambiguity of personal/common name is manifested only in their single use. If a family name follows, it becomes the instrument of the full disambiguation of the personal name.

Another specific feature of the Bulgarian name system is the official name configuration of three names – personal, father’s name and family (the last two morphologically calculable). Although used mainly in official documents, that configuration has to be included in the rules for proper name identification. Moreover it doesn’t hamper but facilitates the identification – the first unit is matched and other two – calculated.

The use of the morphological rules - a specific tuning of the English NE grammars to the Slavonic languages, reduces significantly the volume of the names lists. The already mentioned 330 000 records are identified through lists of 6 500 personal and 2475 foreign family names.

2.2 Organization names
The differences in the name-formation mechanism for the proper names and the names of institutions, organizations, companies etc. is not only in the speed of their updating. The organization names become out of date very soon, they change the rules of their formation continuously.

Here, like the proper names identification, three types of extraction can be set up: 1. Pattern matching in a list (of accompanying elements or candidates for constituents); 2. The morphological calculation; 3. The specific pattern on the alphabet and punctuation level.

The organization names represent a quite complex and freely compiled nominal structure in the most cases. For this reason the organization name can’t be identified on the alphabet and punctuation level only – the 3rd case. The calculation of their grammatical structure, the 2nd case, because of their complexity, can be used only after the tagging or the parsing – procedures missing on this stage of recognition.

Hence, only the first method is applicable – their identification through matching in lists containing

- accompanying elements in their neighborhood
- candidates elements (their potential constituents)

At this stage of the experiment we don’t deal with the latter list – of constituents. Its use is transferred to the next stages of the analysis and the experimental processing of the whole text base.

The pattern matching and the recognition of the separate components of the complex name unit is hindered by the diversity of grammar shape of nominal groups in Slavonic languages. There the syntactic function of the nominal complex and the syntactic links between its components redefine different grammatical values and different flexions for their expression. This variety renders direct matching unfeasible and makes obvious the need of a tagging procedure. It will link the grammatical forms in the complex unit to their invariants given in the list. This obligatory transfer of the recognition of the complex name through its component to the next stages of the analysis is specific for Slavonic language with their rich inflectional system. See, e.g. the variants of the grammatical shape of the components in an English name of an organization - some minor changes in used stop-words as articles and prepositions, easily recognized.

In this way the basic instrument of the recognition of the names of organization remains the lists of their accompanying elements – words and signs.
Elements accompanying and identifying the names of institutions

1. Accompanying words
Bulgarian administrative and company law has not so longstanding traditions as the English one. For this reason not only as a rule, but as usage also, the diagnostic elements in postposition like English LTD, Inc (Bulgarian ООД, ЕАД etc.) occur primarily in official documents but not so often in newspaper narrative texts.

2. Accompanying signs.
Bulgarian orthography rules point to the quotation as the main denouncing instrument. Hence the need of investigation of the quoted expressions follows. This investigation can be done only on a large text base as the quotation of a given expression is an element of usage and not of norm.

This instrument of quoting however is a more out-dated printing standard. In Bulgarian publishing practice, especially in the newspapers, the modern tools for highlighting the notated elements – e.g. font formatting – italics etc. are missing. They are used mostly in some modern editions.

Thus in a certain way the old fashioned publishing standard facilitates one of the most modern application in NLP – the information extraction.

Before the investigation of the text base, a short inquiry in the orthography shows that the quotations marks are used not only to denote named entities but also for citations and stylistic marks - irony, stressing etc. That rule reminds us that the quoting marks are ambiguous in their function – denomination and citation (with a variety of functions). The extraction of quoted names is in fact a disambiguation of the quoting marks.

Types of quoted strings in the text base

Two text collections were examined:
♦ Bulgarian newspapers, volume of 150 000 words, obtained from the WEB-version of the newspaper “Monitor” January-march 2001;
♦ Russian newspapers, 30 000 words, obtained from the CD edition of the newspaper “Nezavisimaja gazeta”, 1998
The two collections differ not only in volume – that difference will be balanced in the next experiments, but in the style and character of publishing and from here in their reliability to compile the rules.

Bulgarian text base is an electronic edition, semi-official, disseminated as WEB document before the final issue. It is more an instrument for collecting readers’ opinions and comments. Its text is not edited as thoroughly the paper edition and the percentage of errors is quite high.

In the Russian base we have the opposite situation. – the CD edition is made after the final edition, the materials are classified by topics, it is extra edited and checked. Thus follows the difference in the behaviour of the quoted expressions that will be discussed below.

**Disambiguation of the quoted expressions**

The observations on the quoted expressions in the two text bases, Bulgarian and Russian, indicate the following formal markers of their function that were used in the compilation of the disambiguating rules (name or citation).

♦ The quoted expression beginning with lower case, is a *citation*.
♦ The quoted expression beginning with uppercase is a *name* if only sentence boundary markers are not included in it. In the latter case it is a *citation* containing sentence or other big text segment.

The distribution of the quoted expressions in both bases, disambiguated by these rules, showed that:
♦ In the Bulgarian corpus we have 365 quoted names and 65 citations.
♦ In the Russian corpus, 5 times smaller, we have 61 quoted names and only 6 citations.

It is to note that the proportionality of the distribution of the quoted names and their volume (6:1) is obvious but there is a great difference in the distribution of citations (11:1). The latter difference can be due to the already mentioned difference in the reliability of the both bases as error checked but can be due to the general difference in genre and style of these editions. The Bulgarian edition is a more “yellow” and the Russian belongs to the more serious journalism – the frequent citation might be a sign of a “lighter” style.

2.3 Dates
The dates and their recording are maybe the only type of NE where the English rules are not extended and enriched in their Slavonic tuning, but on the contrary, they are simplified especially in the official writing. The reason is the writing tradition of the both languages again.

Official writing of dates

The formats of recording time intervals in English are obviously too various. Bulgarian and Russian do not have the am/pm distinction, they don’t use the “/” as separator of year, month and day. There is no tradition to write the day of the week as a part of the date – that is more facultative, in narrative texts.

The standard separator of the time intervals is the point. Semicolon is not in use. Roman digits are used for months. The abbreviations of month names are not used. For these reasons the English rules of dates identification are reduced. The lists containing data components are shorter.

Dates in a narrative text.

The simple formats for date recording in their official writing is compensated by the variety of their writing in a narrative text.

Here we can note that the linguistic tools for the expression of the fixed temporal moment in Slavonic languages are complicated by the morphological instruments needed for the matching of their pattern. E.g. English 1st is allways 1st in every syntactic function but Slavonic 1-ви /1-ый can have 9 flexions for Bulgarian and 18 flexions for Russian following the grammatical values of the noun denoting the time period.

3. Slavonic named entities – text behavior and experiments – further work

The above observations on the linguistic nature of Slavonic NE are based only on their general characteristics and on the general conclusions on their behavior in the text.
However, the real behavior of these units in the text, can be examined only after the real application of the newly-compiled JAPE rules on the large text corpora from both languages - at first on the equal volume – the already mentioned 150 000 words for both languages.

The evaluation process consists of two parts: corpus annotation and performance evaluation, which can both be done within GATE. The corpus annotation will be done semi-automatically by running the Slavonic NE modules over the corpus and then correcting/adding new annotations manually. Performance evaluation will be carried out using the evaluation tool (AnnotationDiff), which enables automated performance measurement and visualisation of the results, and the benchmarking tool, which enables the tracking of a system's progress and regression testing [Cun02b].

These evaluation experiments are planned for the next stage of the work.

References:


