Generating a dictionary of control models for event extraction

Fedor Nikolaev
fsqcds@gmail.com

Vladimir Ivanov
nomemm@gmail.com

Kazan Federal University

Abstract

A subordination dictionary is important in a number of text processing applications. We present a method for generating such dictionary for Russian verbs using Google Books Ngram data. An intended purpose of the dictionary is an event extraction system for Russian that uses the dictionary to define extraction patterns.

1 Introduction and Motivation

Event extraction is an important task in information extraction from unstructured text. This task attracted number of researcher in last decade. An event extraction system aims at capturing certain parts of a text (e.g. event type, participants and attributes). One of the central concepts in event extraction is a trigger word (usually a separate verb) denoting a type of an event [1]. On one hand, the trigger word indicates presence of an event in a sentence. On the other hand, the trigger is considered as a main part in knowledge-based (KB) approach to event extraction.

According to this approach, rules (or patterns) and dictionaries are used. These patterns may be generated automatically [2] or defined manually [3]. However, in languages with free word order (e.g. Russian) a developer of that patterns should also take into account all possible arrangements of words in a sentence. In this case it is more natural to define pattern parts as independent pairs: "event-participant" which will be automatically mapped to "predicate-argument" pairs that denote a subordination in a parse tree of a sentence at hand. Thus a complete subordination dictionary becomes a crucial element of a knowledge-based event extraction system. A well-known limitation of recent works in this area is insufficient dictionary size that prevents using such dictionaries in a computer system.

In 2013 Klyshinsky et al. [4] generated such dictionary for Russian verbs using a set of web corpora; all corpora together contain about 10-11 billion tokens. Authors proposed a method for automatic generation of dictionary for verbs and prepositions. Klyshinsky et al. reported that the dictionary size was about 25-30 thousand verbs. Their method deals only with lexical information, i.e. extraction of verb(-preposition)-noun dependencies was done with six simple finite automata, and no parsing step was performed. Treebanks of Russian language also have insufficient corpus size for automatic generation of a complete (for most Russian verbs) subordination dictionary. The main difference with previous woks that ambiguous part of text was not processed at all. Resulted set was filtered to exclude case ambiguity, infrequent words and ngrams that are not allowed in Russian grammar. The dictionary was evaluated on a corpora of Russian fiction texts and texts from news site and showed good results.

In this paper we present a alternative method for generating a subordination dictionary using a Google Books Ngram Corpus (contains of 67 billion tokens). Main motivation behind this work is to facilitate an event extraction system for Russian that is focused on event types described in ACE [1]. Here we consider the case when a trigger is the main verb (or predicate) that acts as a syntactic head for all participants of a corresponding event (participants of the event act as syntactical arguments of the predicate). We start with a brief overview of user interface that can be used for both pattern definition and dictionary correction. Then we describe the method for generation a subordination dictionary.

2 User interface for pattern and dictionary construction

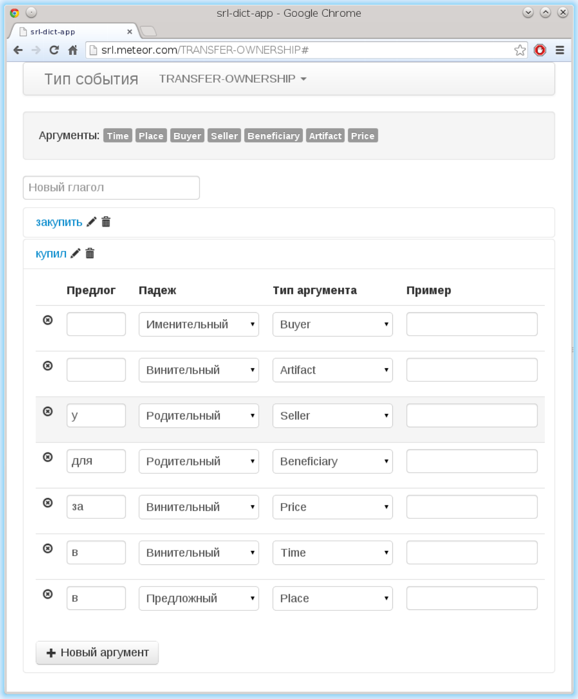
For managing our dictionary we developed a user interface, shown in Figure 1, that allows to define non-linear extraction patterns. A type of the event can be chosen from a drop-down in the top bar. The panel below shows argument types for the event type. There is an interface for dealing with verbs. Existing verbs can be edited and new verbs can be added. In a simple tabular interface user can set preposition, grammatical case of the argument and select participant type. For a few events and triggers using this application for filling dictionary might be enough, but it becomes harder to define all the prepositions and relevant cases as the number of event types and verbs grows.

The method we propose for subordination dictionary generation is based on processing Google Books Ngram data set. The study was carried out for Russian, but this method is applicable to other languages for which Google Books Ngram Corpus and morphological dictionary are available.

3 A subordination dictionary

The main idea is based on using the Google Books Ngram Corpus (GBNC) that was enriched with morphological information and filtered with certain rules.

Proceedings of the Tenth Spring Researcher's Colloquium on Database and Information Systems, Veliky Novgorod, Russia, 2014


srl-dict-app - Google Chrome
srl-dict-app
srl.meteor.com/TRANSFER-OWNERSHIP#
Новый глагол
закупить
+ Новый аргумент

Тип события: TRANSFER-OWNERSHIP

Аргументы: Time Place Buyer Seller Beneficiary Artifact Price

купил

Предлог	Падеж	Тип аргумента	Пример
	Именительный	Buyer	
	Винительный	Artifact	
у	Родительный	Seller	
для	Родительный	Beneficiary	
за	Винительный	Price	
в	Винительный	Time	
в	Предложный	Place	

Figure 1: A simple user interface for definition of event extraction patterns

3.1 Google Books Ngram Corpus

Russian subset of Google Books Ngram Corpus contains 67,137,666,353 tokens extracted from 591,310 volumes [6], mostly from past three centuries. The most part of books was drawn from university libraries. Each book was scanned with custom equipment and the text was digitized by means of OCR. Only ngrams that appear over 40 times across the corpus are included to dataset.

3.2 Coprus preprocessing

The original GBNC data set contains statistics on occurrences of n-grams (n=1…5) as well as frequencies of binary dependencies between words. These binary dependencies represent syntactic links between words from Google Books texts. An accuracy of unlabeled attachment for Russian dependency parser reported in [6] is 86.2%.

As GBNC stores all statistics on a year-by-year basis, each datafile contain tab-separated data in the following format: $ngram, year, match\_count, volume\_count$.

We have preprocessed the original data set in a special way. First, for each dependency 2-gram (the same step for each 3-gram), we have collected all its occurrences on the whole data set and summate all "match_count" values since 1900. Aggregated data set consists of pairs (n-gram, count) for n=2, 3. This step also joined n-grams typed in different cases (lower and upper) into a single (lower case) n-gram.

The next step was to assign each word in a data set a POS-tag and morphological features. For this purpose we used a morphological dictionary provided by OpenCorpora [5] to generate POS-tag and morphological features for 1-grams only.

Thus we got an enriched dataset that has the following format: $n1, match\_count, pos, lemma, gram$,

where $n1$ is a word from the GBNC 1-gram dataset; pos, $lemma$ and $gram$ stand for POS-tag, lemmatized word form and vector of grammatical features respectively. Ambiguous words have led to several records in the this enriched dataset. For instance,

$n1, match\_count, pos, lemma\_id, gramA$
$n1, match\_count, pos, lemma\_id, gramB$

where ambiguous word $n1$ has two sets of grammatical features: $gramA$ and $gramB$. In all such cases we omit these conflicting rows from the dataset, because taking these records into account adds a lot of noise.

3.3 Dictionary of verbal models construction

Let us briefly describe a technique we use for generating a dictionary of direct subject control. To this end we capture all pairs (head, dep) with POS-tag of the head part equals to 'VERB' and having a certain grammatical case for the dependent part (dep), say 'gent' for Genitive. Finally, we group all these pairs by "lemma_id" (in order to regard different forms of the same verb) and count the number of records and summate match_count values. Basically, we run the following SQL-query against the preprocessed dataset:

```
CREATE TABLE direct_verbal_control as
  SELECT
    dep_bigrams.lemma_id,
    dep_bigrams.n1,
  SUM(CASE
    WHEN dep_bigrams.gram LIKE '%nomn%'
    THEN dep_bigrams.count
    ELSE 0 END) AS nomn,
  ...
  SUM(CASE
    WHEN dep_bigrams.gram LIKE '%loct%'
    THEN dep_bigrams.count
    ELSE 0 END) AS loct,
  FROM dep_bigrams
  WHERE dep_bigrams.pos='VERB'
  GROUP BY dep_bigrams.lemma_id;
```

In this example we have six aggregation (sum) functions (one for each grammatical case, e.g. 'loct' for the Locative). Each aggregation function in the query calculates total amount of dependency links between verbs given a lemma_id and arbitrary word forms in a certain grammatical case. We apply the same technique when generating model for control of a preposition from a 3-gram dataset. Queries differ only in the WHEN-condition and GROUP-BY operator that include additional restriction on the second word in a 3-gram.

4 Results and future work

We run two types of queries described in previous section against the whole Google Books Ngram dataset. We have got about 24 thousands of rows (one row per verb) from the dataset of dependency pairs and about 51.5 thousands of rows from the dataset of 3-grams (a verb + preposition per row). Samples from the resulted dictionary are provided in Table 1 and Table 2. The interesting result that

Table 1: A part of generated dictionary for few frequent Russian verbs

Verb	Main case	Genitive	Dative	Accusative	Ablative or Instrumental
сказать	Dat.	0.183	0.573	0.057	0.133
дать	Dat.	0.194	0.511	0.252	0.025
говорить	Dat.	0.192	0.434	0.070	0.166
писать	Dat.	0.207	0.389	0.174	0.123
указать	Dat.	0.216	0.377	0.338	0.056
изменить	Acc.	0.131	0.338	0.352	0.115
объяснить	Ablt. or Instr.	0.093	0.292	0.113	0.489
читать	Acc.	0.196	0.198	0.449	0.102

Table 2: Control of prepositions for verb “купить” (to buy)

Verb	Prep.	Main case	Genitive	Dative	Accusative	Ablt. or Instr.	Locative
купить	для	Gent.	1.0	0.0	0.0	0.0	0.0
купить	из	Gent.	1.0	0.0	0.0	0.0	0.0
купить	без	Gent.	1.0	0.0	0.0	0.0	0.0
купить	до	Gent.	1.0	0.0	0.0	0.0	0.0
купить	с	Gent.	0.595	0.0	0.0	0.405	0.0
купить	в	Loc.	0.0	0.011	0.068	0.0	0.921
купить	за	Ablt. or Instr.	0.0	0.0	0.393	0.607	0.0
купить	к	Dat.	0.0	1.0	0.0	0.0	0.0
купить	на	Loc.	0.0	0.049	0.138	0.005	0.808
купить	по	Dat.	0.0	1.0	0.0	0.0	0.0
купить	под	Ablt. or Instr.	0.0	0.0	0.0	1.0	0.0
купить	со	Ablt. or Instr.	0.0	0.0	0.0	1.0	0.0

many verbs can subordinate words in almost any grammatical case. This result differs significantly from the results presented in [4]. We cannot consider this as an error of our calculation or the parsing method, but rather as an effect of variations in sense of the verb. It might be useful to compare our dictionary to the dictionary generated from a web corpus [4].

In our future work we will evaluate quality of the obtained dictionary. Finally, the will use the dictionary for definition a set of pattern parts (pairs) in our knowledge-based event extraction system. Those pairs will be marked with event participants manually.

References

[1] George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. The automatic content extraction (ace) program - tasks, data, and evaluation. In *LREC*, 2004.

[2] Daria Dzendzik and Sergey Serebryakov. Semi-automatic generation of linear event extraction patterns for free texts. In Natalia Vassilieva, Denis Turdakov, and Vladimir Ivanov, editors, *SYRCoDIS*, volume 1031 of *CEUR Workshop Proceedings*, pages 5–9. CEUR-WS.org, 2013.

[3] Valery Solovyev, Vladimir Ivanov, Rinat Gareev, Sergey Serebryakov, and Natalia Vassilieva. Methodology for building extraction templates for russian language in knowledge-based ie systems. 2012.

[4] Kochetkova N. A. Klyshinsky E. S. Method of automatic generating of russian verb control models. In *XII National conference of artificial intelligence*, 2013. In Russian.

[5] Granovsky D. V. Protopopova E. V. Stepanova M. E. Surikov A. V. Bocharov V. V., Alexeeva S. V. Crowdsourcing morphological annotation. In *Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference “Dialogue” (2013)*, Dialog ’13, 2013.

[6] Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, ACL ’12, pages 169–174, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.