Building Support Tools for Russian-Language Information Extraction

Mian Du, Peter von Etter, Mikhail Kopotev,
Mikhail Novikov, Natalia Tarbeeva, and Roman Yangarber

Department of Computer Science,
University of Helsinki, Finland
{mian.du,peter.etter,mikhail.kopotev,mikhail.novikov,
natalia.tarbeeva,roman.yangarber}@cs.helsinki.fi

Abstract. There is currently a paucity of publicly available NLP tools to support analysis of Russian-language text. This especially concerns higher-level applications, such as Information Extraction. We present work on tools for information extraction from text in Russian in the domain of on-line news. On the lower level we employ the AOT toolkit for natural language processing, which provides modules for morphological analysis and partial syntactic chunking. Since the outputs of both lower-level modules contain unresolved ambiguity, we synthesize the outputs and pass the result into a pre-existing English-language analysis pipeline. We describe how the information extraction system is adapted for multi-lingual support, including extensions to the ontologies and to the pattern matching mechanism. While this is work in progress, we present an end-to-end pipeline for event extraction from Russian-language news.

1 Introduction

We describe work in an on-going project, to adapt the PULS information extraction (IE) system, [3,2] to extend an existing English-language IE system, to extract structured content from Russian-language on-line news. While the IE system has been applied to many different news domains, in this paper we focus specifically on the *border-security* domain.

The English-language IE system contains modules for morphological, shallow–syntactic and semantic analysis. For Russian, the system requires analogous components. Building morphological and syntactic analyzers from scratch is infeasible in a short time-span, due to the immense complexity of the language; therefore we tried to use existing tools for this purpose. However, at present, there is a dearth of publicly available tools for Russian-language natural language processing (NLP). This especially concerns higher-level applications, such as Information Extraction (IE), but is true as well of tools for lower-level analysis. For example, it is reported that the University of Sheffield GATE system, [4], which supports multi-lingual IE, has been adapted to Russian as part of the MUSE-3 project, but there is little information available on its functionality.

After a thorough evaluation of the available linguistic resources for Russian, [1], we chose the AOT toolkit, (www.aot.ru), as the most promising of the existing freely-available tools. The situation with resources is somewhat better; for example, we

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 380–387, 2011.

incorporated a comprehensive geographic gazetteer available as part of the multi-lingual GeoNames database, (www.geonames.org). In this paper, we describe the current status of the project, the components integrated so far, and outline next steps for building a system for analysis of Russian text.

2 Background and Context of the Work

We now briefly describe a specific context in which the Russian-language analysis tools are being applied, which serves as a motivating case-study.

At present various government authorities acknowledge that a significant amount of information useful for monitoring situations relating to public safety is publicly available in the form of published material on the Internet. This has led to an interest in advanced tools that combine techniques from text mining, machine learning, statistical analysis and computational linguistics to help analysts and intelligence experts to manage the growing volume of information, to filter out the irrelevant material, and to extract valuable knowledge from on-line sources.

We collaborate with partners, including the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, to facilitate the process of extracting structured information on events related to border security from on-line news articles. A particular focus is placed on incidents of illegal migration, cross-border criminal activity, and crisis situations at the EU borders. The rationale for exploiting on-line media sources for this purpose is threefold. First, information on certain border security-related events might not be available from other sources, or it might be incomplete in those sources. Second, such information might be available from other sources (e.g., through dedicated networks), but there might be significant delays before the information passes through official channels. Third, open-source information on-line media, can be used for cross-verification against information obtained from other sources.

The need for strengthening the capabilities for tracking the security situation for illegal migration has been identified and acknowledged by the European Commission (EC)[1]. Specifically, the Commission Communication COM (2008) 68A proposes the creation of an Integrated European Border Surveillance System (EUROSUR), and suggests development and deployment of new tools for strategic information to be gathered from various sources (e.g., from open sources) in order to recognize patterns and analyze trends, supporting the analysis of migration routes and the prediction of risks.

The specifications of the EUROSUR network outlined above impose certain requirements on the tools for on-line news event extraction, in particular, they should: extract information in real or near-real time, extract as fine-grained event descriptions as possible, process news articles in many different languages, since a large fraction of relevant events are only reported in non-English, local news.

In this setting, special emphasis is placed on multi-lingual processing, since information about various geographic areas is typically published in different languages,

[1] *Examining the creation of EUROSUR*, `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2008:0068:FIN:EN:PDF`.

and for the system to be useful and to have sufficient coverage, application to multiple languages is an important requirement.

To address these needs, we undertook the work described in this paper.

3 Baseline English System

As a basis, we use the PULS IE system, described in, e.g., [12]. PULS is similar in design in many aspects to GATE, and it had been adapted to various domains; however, to date the support in PULS has mainly focused on English-language text [7,5]. PULS contains modules for lower-level (morphological and syntactic) as well as higher-level (semantic) analysis, and at the end of the pipeline produces filled *templates* extracted from an input corpus. An output template is a *structured* description of a real-world event, in the subject domain. For example in the border-security domain, a news article about smuggling of illegal materials between two countries should induce an event of type "*smuggling*", to which the system should attach the names of the locations/countries involved, the date of the incident, the perpetrators, the type and amount of goods smuggled, etc.

The link from syntactic to semantic analysis is provided by pattern matching: a pattern is a sequence of elements to be matched on input text. The elements can be stated in terms of syntactic, morphological, as well as semantic constraints. A matched pattern invokes an *action*, which can group matched elements into higher-level objects, and eventually into events; e.g., a noun phrase is composed of nouns with modifying adjectives and numerals—*four illegal immigrants*, etc.

Ontologies and Concepts: in defining IE patterns, it is common to group concepts into ontologies, to improve coverage. For example, the concept "contraband" would include the sub-classes "drug", "weapon" and "animal", each of which, in turn contains many sub-concepts; then a single pattern can be stated in terms of the high-level concept, to capture all kinds of contraband.

Inference Rules: Patterns match sequences of words in a sentence. A higher-level mechanism for detecting events is *inference rules*, similar to those employed in expert systems. The job of a pattern is to transform "syntactic" objects—words in the sentence—into "logical" or semantic objects. Inference rules operate strictly at the logical level, at a higher level of abstraction. For example, an inference rule may state that if a smuggling event is mentioned in the text, and there is a mention of known drug or weapon within one sentence from the mention of the event, we can (probably) assume that it refers to the smuggled items. We implemented an inference rule module, based on one described in [14].

4 AOT

The AOT project ("automated processing of text" in Russian) grew out of the DIALING project, [10], which was a commercial project on automatic translation, ended in 2001. Subsequently, components that were used for linguistic analysis were transformed into the AOT toolkit, [9], released under the open-source GNU LGPL license. AOT is a

Table 1. Output of Lemm, the AOT morphological analyzer, adapted with English labels (morphological ambiguity preserved)

Byte	Surface	Lemma	POS	Morphological tags
0	На	на	Prep	—
3	берегу	беречь	Finverb	Impf Transv Act Pres 1p Sg
3	берегу	берег	Noun	Inan Masc Sg {Dat|Loc}
10	пограничной	пограничный	Adj	Fem Sg Anim Inan {Gen|Acc|Inst|Loc}
22	реки	река	Noun	Inan Fem {Sg Gen|Pl Nom|Pl Acc}
27	задержано	задержать	SParticip	Perf Transv Anim Inan Past Pass Sg Neut
36	двадцать	двадцать	Card	{Nom|Acc}
45	семь	семь	Card	{Nom|Acc}
50	нелегалов	нелегал	Noun	Anim Masc Pl {Gen|Acc}

Table 2. Output of Synan, the AOT shallow syntactic analyzer, adapted with English labels

Relations

Type	Parent			Child		
	ID	Surface	Lemma	ID	Surface	Lemma
Num-Noun	7	нелегалов	НЕЛЕГАЛ	5	двадцать семь	—
Adj-Noun	3	реки	РЕКА	2	пограничной	ПОГРАНИЧНЫЙ
Gen-Nom-Group	1	берегу	БЕРЕГ	3	реки	РЕКА
Prep-Group	0	На	НА	1	берегу	БЕРЕГ

Groups

Type	Members
Cardinal-Ordinal-Group	двадцать(5) семь(6)

collection of modules for natural language processing, including libraries for morphological, syntactic, and semantic analysis, language generation, tools for working with dictionaries, and GUIs for visualizing the analysis. In work described below, we use only the morphological and syntactic analyzers, called *Lemm* and *Synan*. (The module for semantic analysis appears to be unfinished, [11].) These analyzers needed to be adapted for our purpose, to correct certain inaccuracies and to output more information, as they were originally designed for different purposes. A major complicating factor resulting from the evolution of the project is incomplete documentation.

In the examples that follow, we use a simple input sentence На берегу пограничной реки задержано двадцать семь нелегалов , ("twenty seven illegal migrants have been detained on the bank of the borderline river").

Lemm Morphological Analyzer: The output of Lemm's morphological analysis is shown in table 1. The columns indicate the byte offset of the word, the surface form, the lemma (or base form), the part of speech, and the morphological tags. As appropriate for a pure morphological analyzer, Lemm does not attempt to resolve ambiguity, and passes it downstream. For example, the surface form берегу, derives from the lemma for the noun "(river) bank", but may also be an inflection of the verb "preserve." Finer ambiguity, on the level of morphological tags, is indicated by |, as when the case is ambiguous for a given lemma.

Synan Syntactic Analyzer: Synan attempts to generate a complete syntactic dependency parse tree; in general it produces a collection of tree fragments. Synan output for the sample sentence is shown in table 2. Synan identifies binary parent-child relations, and "groups"; a group is a sequence of words which function syntactically as an atom, and are not analyzed for dependency (e.g., "twenty seven"). In the process, it resolves morphological ambiguity.

5 System Integration

For building the Russian-language IE system, we integrated the following phases:

- Identify and pre-categorize Russian documents
- Linguistic analysis: morphology, syntax, semantics
- Filling event slots

The input documents are gathered by a dedicated Web crawler, [8], which harvests news articles in Russian matching a large list of keywords that are indicators of potential relevance for the target domain—in this paper, cross-border crime, such as drug smuggling or human trafficking.

5.1 AOT Wrapper

The document text is next processed by the AOT tools. Neither Lemm nor Synan alone extract sufficient syntactic information from the text for building patterns: Lemm, because it does not resolve ambiguity, and Synan, because it does not process all words, only those that participate in recognized relations/groups. Thus we wrapped these two modules into a single, combined analyzer. The purpose of the wrapper is to output a complete analysis of all the words in the sentence, with as much ambiguity removed as possible, and as many relations identified as possible. The wrapper goes through the following stages.

Pre-processing of Lemm and Synan Output: We parse the XML-like outputs of Lemm and Synan into structure shown above, in tables 1 and 2. The part-of-speech (POS) tags and morphological and syntactic tags are mapped into common English tags, (as shown), from the original AOT-specific encoding.

We then normalize the groups to look like other binary relations (by chaining the words in the group), and correct certain inconsistencies or inaccuracies that we have identified in AOT; for example, the analysis of patronymics in Russian names was not appropriate for our purposes in the original Lemm output.

Unifying Synan and Lemm: For every binary relation in Synan output, we take the corresponding parent and child analyses and find corresponding roles in the Lemm output, removing all other analyses. If the lemma for parent or child was null—as, e.g., when the corresponding element was a group—we infer information from Lemm output for the missing element. In cases when a word does not participate in any relation identified by Synan, its analysis is taken entirely from Lemm output, passing along any unresolved ambiguity.

Table 3. Result of the PULS AOT wrapper, combining morphology and syntax, and resolving ambiguity

ID	Offset	Surface	Lemma	Relation	POS	Morphological tags
0	0	На	на	—	prep	
1	3	берегу	берег	prep-group→0	noun	2genl inan masc loc sg
2	10	пограничной	пограничный	adj-noun→3	adj	inan anim fem gen sg
3	22	реки	река	gen-nom-group→1	noun	inan fem gen sg
4	27	задержано	задержать	—	sparticip	past pass sg neut perf transv
5	37	двадцать	двадцать	card-ord-group→6	card	nom
6	46	семь	семь	num-noun→7	card	nom
7	51	нелегалов	нелегал	—	noun	anim masc gen pl

Output Generation: After the unification, we assemble the resulting analyses for all words into a parse tree, or into a set of tree fragments. In some cases involving conjunctions, AOT produces two parents for a node; the wrapper adjusts the links so that they form a proper tree structure.

Table 3 shows the wrapper's output, which was modeled after Connexor parser output, [6]. This serves as the basis for the subsequent semantic analysis and IE.

5.2 Information Extraction

Adapting the Ontology for Russian: We use the existing PULS English-language ontology—including the domain-specific concepts—as the shared or *interlingua* concept base, and link directly to these concepts their instances in Russian (and other languages in the future). The base ontology needed to be extended in some cases, e.g., by making explicit certain concepts that may be ambiguous in English. For example, an English word (e.g., "arrest") can act as verb and noun, whereas in Russian they have different base forms—"арестовать" vs. "арест".

Patterns and Inference Rules: Patterns are used to group smaller syntactic units together into larger units, starting from the individual words in the sentence syntactically analyzed by AOT. For example, adjectives and numeric expressions are joined with the nouns they modify into noun phrases, genitive groups ("the house of the president of France") are joined into larger noun phrases, then into prepositional phrases, and so on. After the higher-level phrases are built from lower-level elements, domain-specific patterns and inference rules are applied to find events, based on semantic analysis, as follows.

We can construct an inference rule for finding arrest events in a sentence, e.g., of the kind *"perpetrator detained in location"*—for text like "преступник был задержан в Греции". The rule should fire if it finds phrases headed by concepts of type *perpetrator*, *arrest* and a locative prepositional phrase with *location*; further, the rule may specify that the concepts should be linked by a certain syntactic relation, or (more loosely) occur in the same sentence, or in nearby sentences. When the rule fires, it specifies which slots in the event are filled by which semantic constituents. An event template (partially) filled by such a rule is shown in Fig. 1, on a real-world news article. More advanced rules and patterns can exploit additional morphological and syntactic constraints.

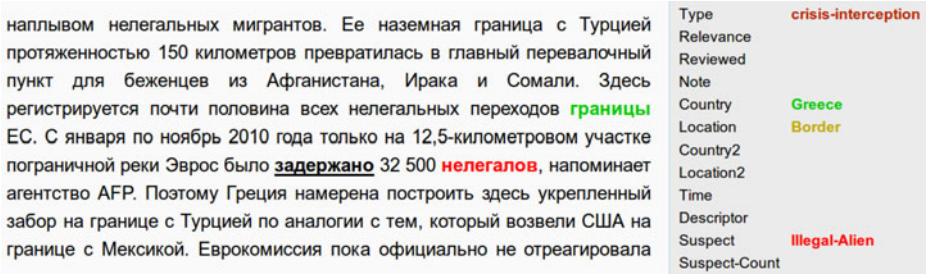

Fig. 1. A sample document text, and a (partially) filled-in event template

6 Current Work

The framework we describe provides end-to-end functionality for Russian IE. Our current efforts center on improving performance, which mainly entails enriching the concept ontology and building patterns and rules. We pursue this via two approaches in parallel. First, we are adapting the patterns and rules in the pre-existing English-language system to Russian text, which involves a certain amount of manual labor. Second, this is aided by *bootstrapping* for lexicons and patterns, as tested previously on English and other languages, e.g., [15,13], which we are adapting for Russian. Lastly, further work is required on the AOT wrapper, as well as on extending the mechanisms for IE patterns and rules to utilize the full range of morphological and syntactic information provided by the lower-level analysis, which will help improve precision.

References

1. Astaf'eva, I., Bonch-Osmolovskaya, A., Garejshina, A., Grishina, J., D'jachkov, V., Ionov, M., Koroleva, A., Kudrinsky, M., Lityagina, A., Luchina, E., Sidorova, E., Toldova, S., Lyashevskaya, O.S.S., Koval', S.: NLP evaluation: Russian morphological parsers. In: Proceedings of Dialog Conference, Moscow, Russia (2010)
2. Atkinson, M., Belyaeva, J., Zavarella, V., Piskorski, J., Huttunen, S., Vihavainen, A., Yangarber, R.: News mining for border security intelligence. In: Proceedings of IEEE ISI-2010: Intelligence and Security Informatics, Vancouver, BC, Canada (2010)
3. Atkinson, M., Piskorski, J., Tanev, H., van der Goot, E., Yangarber, R., Zavarella, V.: Automated event extraction in the domain of border security. In: Proceedings of MINUCS: Workshop on Mining User-Generated Content for Security, at the UCMedia: 1st International ICST Conference on User-Centric Media, Venice, Italy (2009)
4. Bontcheva, K., Maynard, D., Tablan, V., Cunningham, H.: GATE: A Unicode-based infrastructure supporting multilingual information extraction. In: Proceedings of Workshop on Information Extraction for Slavonic and other Central and Eastern European Languages, Borovets, Bulgaria (2003)
5. von Etter, P., Huttunen, S., Vihavainen, A., Vuorinen, M., Yangarber, R.: Assessment of utility in Web mining for the domain of public health. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pp. 29–37. Association for Computational Linguistics, Los Angeles (June 2010), http://www.aclweb.org/anthology/W10-1105

6. Järvinen, T., Tapanainen, P.: A dependency parser for English. Tech. Rep. TR-1, Department of General Linguistics, University of Helsinki, Finland (February 1997)
7. Linge, J., Steinberger, R., Weber, T., Yangarber, R., van der Goot, E., Khudhairy, D.A., Stilianakis, N.: Internet surveillance systems for early alerting of health threats. Eurosurveillance Journal 14(13) (2009)
8. Piskorski, J., Atkinson, M., Belyaeva, J., Zavarella, V., Huttunen, S., Yangarber, R.: Real-time text mining in multilingual news for the creation of a pre-frontier intelligence picture. In: Proceedings of ISI-KDD: ACM SIGKDD Workshop on Intelligence and Security Informatics, at KDD-2010: 16th Conference on Knowledge Discovery and Data Mining, Washington, DC (2010)
9. Sokirko, A.: Semantic dictionaries in automatic text analysis, based on DIALING system materials. Ph.D. thesis, Russian State University for the Humanities, Moscow (2001)
10. Sokirko, A.: A short description of DIALING project (2001), `http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html`
11. Sokirko, A.: Private communication (2011)
12. Steinberger, R., Fuart, F., van der Goot, E., Best, C., von Etter, P., Yangarber, R.: Text mining from the web for medical intelligence. In: Perrotta, D., Piskorski, J., Soulié-Fogelman, F., Steinberger, R. (eds.) Mining Massive Data Sets for Security, OIS Press, Amsterdam (2008)
13. Thelen, M., Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002 (2002)
14. Wilensky, R.: Common LISPcraft. W. W. Norton and Company, USA (1986)
15. Yangarber, R.: Counter-training in discovery of semantic patterns. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan (July 2003)