Information Extraction from the Weather Reports in Serbian

Staša Vujičić Stanković
University of Belgrade
Faculty of Mathematics
Studentski trg 16, Belgrade
Serbia
+381 11 202 78 01
stasa@matf.bg.ac.rs

Vesna Pajić
University of Belgrade
Faculty of Agriculture
Nemanjina 6, Zemun
Serbia
+381 64 2977630
svesna@agrif.bg.ac.rs

ABSTRACT

In this paper, we describe a process of extracting information from meteorological texts in Serbian. The text corpus consists of almost 46000 sentences. Having in mind the specifics of Serbian and characteristics of meteorological sublanguage, we develop a classification schema for structuring extracted information and transducers for annotating pieces of information in the text corpus. We describe the transducer for extracting information about daily temperatures and give some evaluation parameters for all other transducers used in the information extraction process.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis, Language parsing and understanding*; H.3.0 [**Information Storage and Retrieval**]: General

General Terms

Algorithms, Experimentation, Languages, Performance

Keywords

Information extraction, transducers, Serbian language, language resources

1. INTRODUCTION

Weather forecast reports are interesting for natural language processing because of their properties and the possibility of various uses of extracted data. These texts have been studied over the years in the areas such as information extraction, text mining or text understanding, and the obtained data were used for machine translation from one language to another (TAUM-METEO system developed in Canada for machine translation from English to French and vice versa [2] and [14]), data visualization described in [5], web information extraction using extraction ontologies represented in [11], creating dialogue manager system as in [1], summarization of data from multiple sources ([6] and [7]), etc.

In this paper we present the process of extracting information about weather conditions from meteorological texts in Serbian, which can be used for different purposes (for example, for automatic creation of lexicon or annotation of texts). The main goal of this research was to provide foundations for developing electronic resources in Serbian, construction of sublanguages, ontologies, machine translation system from Serbian to English, and vice versa, and different kinds of linguistic researches in the domain of weather forecast. Some specifics of Serbian that are important for this research are presented in Section 2. The corpus of meteorological texts in Serbian, collected during 2010, 2011, and 2012 years from several sources is described in Section 3.

The main goal of the extraction process was to annotate information contained in a text description. Three types of information were of interest: location, time, and meteorological phenomena. Semantic classes of information used to structure the data are described in Section 4.

The process of information extraction is presented in Section 5. The extraction rules are defined by finite state transducers (FST) ([4] and [15]) and recursive transition networks (RTN) with output ([4] and [16]), both referred to as transducers in this paper. We used the UNITEX software system [12] for the creation and application of the transducers.

Finally, we evaluate the information extraction process and give the directions for the future research.

2. THE SPECIFICS OF SERBIAN

The specific features of Serbian determine, to a great extent, approach and method that will be used for the information extraction from texts written in Serbian.

Serbian is a language with rich morphology. For example, the most adjectives in Serbian may take more than 40 different forms. There are algorithms for different Natural Language Processing (NLP) tasks that have excellent results when applied to texts in English, but very bad when it comes to texts in a language with rich morphology, such as Serbian. The rich morphological system of Serbian requires the use of additional linguistic resources, such as electronic dictionaries and grammars, for text processing. Thus, it is possible to develop systems for the information extraction that would be efficient when applied to texts in Serbian.

This paper describes a process of extracting information from texts in Serbian, in which the electronic dictionary for Serbian ([8] and [9]) was used. This dictionary was written in the DELA format [13]. It contains 125269 lemmas of simple words and 4378245 simple word forms, as well as 5251 lemmas of compounds and 106731 forms of compounds [10].

3. THE CHARACTERISTICS OF THE TEXT CORPUS

Meteorological texts have been collected during 2010, 2011, and 2012 years from several sources (Republic Hydrometeorological

BCI'12, September 16–20, 2012, Novi Sad, Serbia.

Local Proceedings also appeared in ISBN 978-86-7031-200-5, Faculty of Sciences, University of Novi Sad.

Service of Serbia,[1] the Meteos agency,[2] the Politika daily news,[3] B92,[4] SMedia[5] and Internet portal Krstarica[6]). The created text corpus contains 13705 text descriptions, which consist of a total of 45862 sentences.

3.1 Weather Forecast Sublanguage

The language used for describing weather conditions in textual reports is very specific and easily recognizable. A limited set of words from natural language, which is used to describe the meteorological phenomenon, can be treated as a sublanguage, along with its characteristics:

- limited vocabulary – the same words are used to describe a meteorological phenomenon in almost every weather report;
- irregular syntax – sentences in meteorological reports typically do not contain auxiliary verb, and often do not have a predicate ("*Vetar slab*, *jugoistočni*." – "*Wind weak*, *southeast*.") or adverbs;
- text structure – it is not possible to distinguish different statements based only on punctuation, since a sentence often contains multiple statements, and a few sentences sometimes merges into one separated with commas.

On the one hand, the existence of such sublanguage facilitates the text processing, since many syntactic rules are simplified in comparison to natural language. On the other hand, it is contempt of natural language syntax rules that prevents the use of existing electronic grammars, developed and available for a given natural language.

3.2 The Structure of Textual Meteorological Descriptions

The descriptions of weather conditions consist of smaller fragments (sentences and parts of sentences), which carry three types of information (meteorological phenomenon, location and time), combined together in a statement. Therefore, every semantic unit of the text structure (particular statement) can be treated as a triple <*location*, *time*, *phenomenon*>. The ideal information extraction process from the following description in Serbian "*Ujutru i pre podne u nižim delovima grada magla ili sumaglica.*" ("*In the morning and before the noon in the lower parts of the city fog or haze.*") would extracts the following triples:

<"*niži delovi grada*", "*ujutru*", "*magla ili sumaglica*">
<"*niži delovi grada*", "*pre podne*", "*magla ili sumaglica*">
(<"*the lower parts of the city*", "*In the morning*", "*fog or haze*">
<"*the lower parts of the city*", "*before the noon*", "*fog or haze*">)

The statements mutually overlap in the textual descriptions, usually with no clear boundary between two different statements. This semantic structure requires a special approach, semantically oriented, in order to resolve coreferences between different parts. However, the first steps in this process are the detection and isolation of the values of individual features. This paper describes exactly this process, while merging isolated pieces of information and their values into the statements will be the subject of a future research.

4. SEMANTIC CLASSES FOR INFORMATION STRUCTURING

The information contained in the textual descriptions of weather conditions, which were of interest in the research, are grouped into semantic classes of different levels. A semantic class, together with possible additional classification, should be assigned to each separate fragment of the text. Hierarchical classes are shown in Table 1.

Table 1: Class Hierarchy Used to Structure Information Extracted from the Text

Type	Element	Feature	Value examples
Meteo	Padavine (*Precipitation*)	TipPadavina (*PrecipitationType*)	*kiša, sneg ...* *(rain, snow...)*
		ObimPadavina (*Precipitation-Amount*)	*slaba, jaka, ...* *(weak, strong...)*
	Oblačnost (*Cloudiness*)	PrisustvoOblaka (*CloudPresence*)	*sunčano, oblačno* *(sunny, cloudy)*
		ObimOblačnosti (*CloudAmount*)	*promenljivo, potpuno...* *(variable, fully...)*
	Vetar (*Wind*)	PravacVetra (*WindDirection*)	*jugoistočni, severni ...* *(southeast, north….)*
		JačinaVetra (*WindAmount*)	*jak, slab...* *(strong, weak...)*
		BrzinaVetra (*WindSpeed*)	*16 m/s*
	Temperatura (*Temperature*)	Temperatura (*Temperature*)	*12 stepeni, 12 C, dva stepena, ispod nule ...* *(12 degrees, 12 C, two degrees, below zero…)*
		KatTemperature (*Temperature-Category*)	*najviša, jutarnja ...* *(maximum, morning…)*
		OpisTemperature (*Temperature-Description*)	*hladno,toplije, porast ...* *(cold, warmer, rising…)*
	Pojava (*Phenomenon*)	TipPojave (*PhenomenonType*)	*magla, oluja ...* *(fog, storm...)*
Location	Teritorija (*Teritory*)	ImeTeritorije (*TeritoryName*)	*Srbija, Evropa, Beograd ...*
		DeoTeritorije (*TeritoryPart*)	*severoistok, južni delovi* *(northeast, southern parts)*
	Lokalitet (*Locality*)	Lokalitet (*Locality*)	*na planinama, lokalno...* *(in the mountains, localy)*
Time	Dan (*Day*)	Datum (*Date*)	*15. januar* *(January 15th)*
		ImeDana (*DayName*)	*ponedeljak, utorak ..* *(Monday, Tuesday...)*
		DeoDana (*DayPart*)	*ujutru, posle podne* *(in the morning, in the afternoon)*
	Period (*Period*)	Period (*Period*)	*sledeće nedelje, tokom februara* *(next week, during February…)*

The names of the features, given in Table 1, are used for annotating pieces of information in the text.

The annotations had the following syntax:

<Feature>text segment</Feature>

[1] http://www.hidmet.gov.rs
[2] http://www.meteos.rs
[3] http://www.politika.rs
[4] http://www.b92.net
[5] http://www.smedia.rs
[6] http://www.krstarica.com

Hence, the example sentence “*U većem delu zemlje promenljivo oblačno, mestimično slaba kiša, pljuskovi i grmljavina*.” (“*In most of the country variable cloudiness, with areas of light rain, showers, and thunder.*”), should be annotated as follows:

<lokalitet>U većem delu zemlje</ lokalitet>
<obimOblacnosti>promenljivo</obimOblacnosti>
<prisustvoOblaka>oblačno</prisustvoOblaka>,
<lokalitet>mestimično</lokalitet>
<obimPadavina>slaba</obimPadavina>
<tipPadavina>kiša </tipPadavina>,
<tipPojave>pljuskovi</tipPojave> i
<tipPojave>grmljavina</tipPojave>.

(<Locality>In most of the country </Locality>
<CloudAmount>variable</CloudAmount>
<CloudPresence>cloudiness</CloudPresence>,
<Locality>with areas</Locality> of
<PrecipitationAmount>light </PrecipitationAmount>
<PrecipitationType>rain</PrecipitationType>,
<PhenomenonType>showers</PhenomenonType> and
<PhenomenonType>thunder</PhenomenonType>.)

5. INFORMATION EXTRACTION PROCESS

We used transducers (FST and RTN) as extraction rules. The transducer that describes the rule for extracting particular piece of information was created for each feature given in Table 1. The rules were applied through the software system UNITEX, where the structuring of data was done by annotating text segments that carry information. The application of transducers was performed sequentially, one by one. The application order was not important for the majority of created transducers, although it is possible to organize the information extraction process so that the successive application of transducers improves the efficiency of the process (a cascade of transducers, one operating after the other using the results of previously applied transducers [3]). In this section, we will present one of the transducers that extracts information related to the temperature.

Temperature data have been presented in the texts as values (*12 stepeni – 12 degrees, 12°C, 12 C, dva stepena – two degrees, ispod nule – below zero, minus 5 ...*) or descriptive (*hladno - cold, hladnije - colder, toplo - warm, toplije – warmer, pad temperature – the temperature drop, temperatura u porastu - the temperature rising ...*). For each way of representing temperature, a special extraction rule has been created. Figure 1 shows the main transducer (*temperatura.grf*) in the RTN for extracting information related to the temperature.

Subgraph calls are marked with gray colour. Subgraph *vrednost.grf* recognizes different expressions for the specific value (number of degrees) of the temperature. This subgraph is shown in Figure 2.

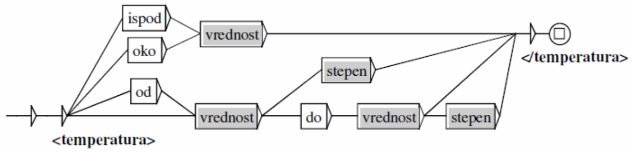

Figure 1: The main transducer *temperatura.grf* within the RTN, for extracting information about the temperature.

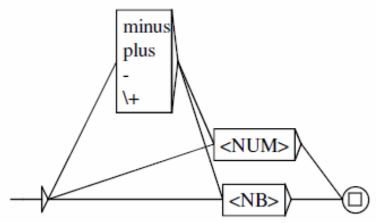

Figure 2: Subgraph *vrednost.grf* that recognizes numeric values written as numbers or text.

The lexical mask <NB> recognizes successive digits. The lexical mask <NUM> recognizes all the words in the dictionary that are marked with a code NUM (*jedan*, *dva*, *tri* - *one*, *two*, *three*, ...). Thus, this subgraph recognizes, among others, the following expressions: *10*, *minus dva – minus two*, *+5* ili *jedanaest – eleven*. The main transducer *temperatura.grf* (Figure 1) contains a subgraph call *stepen.grf*. This graph is intended to recognize expressions that describe the degrees on the Celsius scale, as the common unit of temperature measure, in the texts in Serbian language. Subgraph *stepen.grf* is shown in Figure 3.

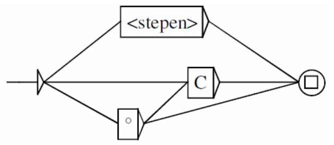

Figure 3: Subgraph *stepen.grf* that recognizes phrases for marking degrees on the Celsius scale.

The lexical mask which refers to a dictionary word (<*stepen*>) recognizes any form of the word *stepen* – degree (*stepena*, *stepeni*, *stepenima* etc.). Graph *temperatura.grf* recognizes the following phrases: *oko +8 °C* (*approximately +8 ° C*), *- 1C*, *- 30 ° C*,*- 4 stepena* (*- 4 degrees*), *od -1 C do 1 C* (*from -1 C to 1 C*), *-1 do +3 stepena* (*-1 to +3 degrees*), *-12 do -8* (*-12 to -8*), *od 11 do 15 stepeni* (*from 11 to 15 degrees*), *11 stepeni* (*11 degrees*), *od pet do devet stepeni* (*from five to nine degrees*), *oko četiri* (*about four*), *ispod 0 (below 0*) etc.

Similarly, for each feature in the Table 1 a rule extraction is created for annotation of the text segments that carry specific information.

5.1 Analysis of Extracted Information and Process Efficiency

The process of information extraction from the meteorological texts is in the initial phase. During this phase, the analysis of the texts from the described corpus was performed and the transducers for extracting simple features were created. Since the extraction rules are still evolving, and the text corpus over which the extraction is carried out is fairly large (45862 sentences with more than one million tokens), a comprehensive evaluation of the system’s efficiency, which would accurately assess the precision and recall, is not currently possible. However, an initial analysis of the created transducers, which would determine the directions for further development, is possible.

Table 2 lists the transducers which were used to extract information, in order of their implementation. The number of extracted text segments is shown in the third column of the table, while the evaluation of precision is presented in the fourth.

Table 2: Performance Evaluation of Graphs Used for the Extraction of Information

Transducer	Features	Number of extracted text segments	Evaluation of precision
opisTemp	*OpisTemperature* (*Temperature-Description*)	11518	100%
temperature	*Temperatura* (*Temperature*)	25618	99.6%
katTemp	*KatTemperature* (*Temperature-Category*)	14817	100%
vetarPre	*JacinaVetra* (*WindAmount*) and *PravacVetra* (*WindDirection*)	7720	100%
vetarPost	*JacinaVetra* (*WindAmount*) and *PravacVetra* (*WindDirection*)	1559	100%
padavine	*TipPadavina* (*PrecipitationType*) and *ObimPadavina* (*Precipitation-Amount*)	18878	100%
oblacnost	*ObimOblacnosti* (*CloudAmount*) and *PrisustvoOblaka* (*CloudPresence*)	18875	98%
deoTeritorije	*DeoTeritorije* (*TeritoryPart*)	4918	99.8%
imeTeritorije	*ImeTeritorije* (*TeritoryName*)	6036	95%
lokalitet	*Lokalitet* (*Locality*)	7623	98%
pojava	*Pojava* (*Phenomenon*)	3737	100%

6. CONCLUSION

The high precision of the transducers is expected, given that this is an early stage of the system design and the extraction rules creation process. Further development of the process, in order to extract a larger number of individual pieces of information (i.e. to increase recall), will surely reduce the precision. However, it is expected the transducers will still maintain high efficiency.

We would like to emphasize that the next step in the process, after the extraction of simple features, is merging the extracted data into classes of higher semantic level. During that process, it will be possible to further improve efficiency, by resolving ambiguities or correcting wrongly interpreted text segments.

7. ACKNOWLEDGMENTS

This research was conducted through the projects 178006 and III 47003, financed by the Serbian Ministry of Science.

8. REFERENCES

[1] Brkić, M., and Matetić, M. 2007. Modeling Natural Language Dialogue for Croatian Weather Forecast System. In *Proceedings of the 18th International Conference on Information and Intelligent Systems* (Varaždin, Croatia, 2003), 391–396.

[2] Chevalier, L., Dansereau, J., and Poulin, G. 1978. *TAUM-METEO: Description du Système*. Universite de Montreal, Canada.

[3] Friburger, N. and Maurel, D. 2004. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science* 313, 1 (2004), 93–104.

[4] Jurafsky, D. and Martin, J. H. 2008. *Speech and language processing*, 2nd edition. Prentice-Hall Inc.

[5] Kerpedjiev, S. and Noncheva, V. 1990. Intelligent Handling of Weather Forecasts. In *Proceedings of the 13th International Conference on Computational Linguistics COLING-90*, 3 (Helsinki, Finland, August 20–25, 1990), 379–381.

[6] Kononenko I., Popov I., and Zagorulko Yu. 1999. Approach to Understanding Weather Forecast Telegrams with Agent-Based Technique. In *Perspectives of System Informatics, Third International Andrei Ershov Memorial Conference, PSI'99* (Novosibirsk, Russia, July 6–9, 1999), 511–516.

[7] Kononenko I., Kononenko S., Popov I., and Zagorulko Yu. 2000. Information extraction from non-segmented text (on the material of weather forecast telegrams). In Proceedings *of the 6th International Conference, RIAO 2000* (College de France, France, April 12–14, 2000), 1069–1088.

[8] Krstev, C. 2008. *Processing of Serbian Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade, Belgrade, Serbia.

[9] Krstev, C. and Vitas, D. 2005. Corpus and Lexicon – Mutual Incompleteness. In *Proceedings from the Corpus Linguistics Conference Series*, 1, 1, ISSN 1747-939 (Birmingham University, UK, July 14–17, 2005).

[10] Krstev, C., Vitas, D., Obradović, I., and Utvić, M. 2011. E-Dictionaries and Finite-State Automata for the Recognition of Named Entities. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing* (Blois, France, July 12–15, 2011), 48–56.

[11] Labsky, M., Nekvasil, M., and Svatek, V. 2007. Towards web information extraction using extraction ontologies and (indirectly) domain ontologies. In *Proceedings of the 4th international conference on Knowledge capture K-CAP '07* (Whistler, BC, Canada, October 28–31, 2007), ACM New York, NY, USA, 201–202.

[12] Paumier, S. 2008. *Unitex 2.1 User Manual*. http://www-igm.univmlv.fr/~unitex/UnitexManual2.1.pdf.

[13] Silberztein, M. 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Edition Masson, Paris.

[14] Slocum, J. 1985. A Survey of Machine Translation: its History, Current Status, and Future Prospects. In*:* *Computational Linguistics* 11, 1 (1985), 1–17.

[15] Vitas, D. 2006. *Prevodioci i interpretatori: Uvod u teoriju i metode kompilacije programskih jezika.* Faculty of Mathematics, University of Belgrade, Belgrade, Serbia.

[16] Woods, W. 1970. Transition network grammars for natural language analysis, In *Communications of the ACM* 13, 10 (1970), 591–606.