TRANSFER OF NATURAL LANGUAGE PROCESSING TECHNOLOGY: EXPERIMENTS, POSSIBILITIES AND LIMITATIONS CASE STUDY : ENGLISH TO SERBIAN

Miroslav Martinović,
College of New Jersey, Computer Science Department

Abstract. In this paper, we describe research and instructional activities related to a scientific experiment. This study was conducted to examine possibilities and limitations of transferring human language technologies from a technology rich and morphologically modest language like English to a technologically developing and morphologically rich language as Serbian. We briefly present some collaborative achievements that include building of an information retrieval system (SPretS), a part-of-speech tagger (SPaS), a shallow parser (SPaSer), and a topic tracking system (SeToTraS) for Serbian language. Some of the results reported here have already been disseminated through conference papers while the other ones are in the process of being so.

(I) Overview

Work described here has been a part of a Sabbatical project envisioned as a range of research activities blended with on-site and/or on-line lecturing activities at a graduate or a special topic undergraduate level.

Research activities clustered around computational linguistic resources and tools, specifically those developed by myself for English language (Anaphora Resolver 'AARLISS' ([35]), Named Entity Recognizer 'HyNER', Lemmatizer/Conflator 'SteLemMin' ([37]), Question Answering System 'QASTIIR' ([34])). Published literature was researched and related tools demoed, analyzed and evaluated. Local researchers were subsequently initiated into conducting a related analogous research in development of CL resources for Serbian.

Lecturing activities included an overview of state of the art methodologies and achievements in the area of Natural Language Processing, with an emphasis on those related to small languages like Serbian. Theory and practice of searching and retrieval of text and bibliographic information were introduced with a special prominence placed on question answering issues. In addition, current and most successful approaches to the problem of identifying the answers to a natural-language question from a large quantity of on-line text documents was studied in depth to set the stage for the concurrent and subsequent research proceedings.

(II) Background

Human language technology field has at long last matured enough and a much greater attention is being paid to developing linguistic resources for languages spoken by less than ten million people (small languas). Many (focused) applications like command-and-control, dictation, voice-mail recognition, audio search, automatic translation of news texts and parliamentary proceedings have developed to reach a high quality for practical use. Natural language systems have been broadened to a wider range of languages. The EU now has 23 official languages and is using a number of machine readable lexicons (MRLs) (e.g. Finnish, Czech, Hungarian, Slovene). US large-scale research projects also targeted MRLs for small languages (e.g. DARPA GALE ([18]), Transtac ([19]): Modern Standard Arabic, Arabic dialects, Farsi, DARPA CAST: Pashto). Numerous new projects/networks are pioneered to develop languages resources (e.g. NEMLAR

([20]): Network for Euro-Mediterranean Language Resources; TELRI ([21]): Trans-European Language Resources Infrastructure focused on Central and Eastern European Languages; LDC's Less Commonly Taught Languages (LCTL) Project ([17]): developing linguistic resources for Urdu, Thai, Hungarian, Bengali, Punjabi, Tamil, Yoruba). New benchmark tasks/evaluations have been introduced to measure progress (Morpho Challenge: addresses evaluation of automatic morphological segmentation/analysis algorithms; CoNLL Shared Task: concentrates on dependency parsing for multiple languages (including Czech, Arabic, Turkish, Basque, Hungarian).

Moreover, ethnolinguistic analysis of the relationship between culture, thought, and language and in particular, a study of a minority language within the context of the majority population has been gaining a significant impetus ([3], [4], [10], [11], [12], [13], [14], [15], [16], [22], [26], [27], [29], [36], [42]). Standardization, linguistic normalization and revitalization of small languages have been initiated and promoted and a rising number of web-pages in lesser-used languages testifies to this fact.

Automatic processing of small languages needs to surmount a number of difficulties which evolve from their special status.

(a) As these languages have few speakers, there are few native linguists and even fewer computational linguists. Traditional rule-based approaches to tagging, parsing, etc. may thus be difficult to apply.

(b) The scarce financial support that these languages enjoy equally seems to virtually exclude rule-based approaches due to the amount of human labor these approaches generally require. This problem might be overcome if computational frameworks derived from other languages can be adopted.

(c) Corpus-based approaches are only applicable if adequate corpora are available. However, creation of a corpus is time- and money-consuming and requires linguistically sound conceptions, especially if general-purpose corpora are to be created.

(d) Example-based approaches appear to be more promising in this light if no general-purpose corpora, but specific examples are required. Compilation of special examples also seems to be easier to implement than to write formal rules. However, little is known of the feasibility of this paradigm with respect to minority languages.

(e) Shallow knowledge techniques may be developed or are already in use, which benefit from a specific property of a language or a language family. This however may hamper the transfer of the approach from one language to other languages. Some techniques might work with analytic languages and not with agglutinative languages, etc. Different writing systems might also prevent one simple approach from being applicable to another language.

(III) Goals and Objectives

The entire field of Computational Linguistics presently lacks in substantive research related to small languages. A number of talented students and researchers from Serbia were, by attending "Information Retrieval" and "Open Domain Question Answering Systems" courses introduced into the state of the art research in the field. Accompanied with a concurrent directed readings and demonstrations of resources and tools developed through corresponding research for English language, this instigated a collaborative work on a number of relevant issues like:

(a) The relation between NLP and small language support in general.

(b) Development of specific NLP applications for Serbian, e.g. tagging, morphological analysis, parsing, information retrieval, question-answering.

(c) Development of corpora and machine-readable dictionaries for Serbian.

(d) Presentation of shallow knowledge NLP techniques which could be applied to Serbian.

(e) Overview studies that describe the state of the art of NLP for Serbian and other small languages of the entire region and the language type.

(f) Comparative analysis of different NLP approaches to different small languages and languages types.

(g) Free resources for NLP, their application areas and limitations.

(h) The requirements for NLP applications for Serbian and related languages.

(IV) Project Plan

During both semesters, a combination of a range of ensuing research activities was blended with on-line or on-site lecturing activities at graduate or a special topic undergraduate level.

During the first semester, instructional activities included a course on Information Retrieval Systems similar to the one taught by myself a number of times. An overview of state of the art methodologies and achievements in the area of Natural Language Processing, with an emphasis on those related to small languages like Serbian was integrated in the course. Theory and practice of searching and retrieval of text and bibliographic information were studied in details with an emphasis on question answering issues. The course was a fusion of lectures, paper presentations, critiques and hands on project design and development. Students took on directed readings and paper presentations and critiques, as well as on the development of an Information Retrieval system of their own as their semester project, and a basis for their future research. This system's architecture was geared towards a flexibility that was hoped to later allow for its transformation into an Information Extraction system, and eventually into a Question Answering system. During this course, students were also introduced to the available linguistic tools for tagging, parsing, reference resolution and named entity recognition, some developed by myself.

The stage was then set for the next semester and the subsequent course on Open Domain Question Answering Systems. Current and most successful approaches to the problem of identifying the answers to a natural-language question from a large quantity of on-line text documents were studied in depth to set the stage for the concurrent and subsequent research proceedings. Similarly to the previous course, this course was a blending of lectures, paper presentations, critiques and hands on project design and development. Students conducted directed readings, and paper presentations and critiques. In addition, they were expected to develop modules later to be integrated into a Question Answering system of their own as their semester project and a basis for the future research. During this course, students also got familiarized with some existing question answering systems, as well as to QAS-TIIR, the one developed by myself.

A collaborative projects stemming from the course projects addressing the specificities of resources and tools for Serbian language were envisioned, promoted, and developed as described later in this paper.

(V) Dissemination and Outcomes

The substantive results of this mission were communicated to the scientific community through four collaborative papers submitted to INTERSPEECH 2007 and CICLING 2008 computational linguistics conferences.

(1) "Building an Information Retrieval System for Serbian – Challenges and Solutions", by Miroslav Martinović, Srđan Vesić, Goran Rakić (submitted, accepted and presented at INTERSPEECH 2007 conference held in Antwerp, in August 2007).

This paper described challenges encountered while building an information retrieval system for Serbian language. As a backbone of our system, SMART retrieval system augmented with features necessary to deal with specificities of the Serbian alphabet was utilized. In addition, morphological richness of the language accentuated implications of the text preprocessing phase. During this phase, two algorithms were devised which increased retrieval precision by 14% and 27%, respectively. Testing was conducted using two gigabyte EBART collection of Serbian newspaper articles.

The overall structure of Serbian Retrieval System (SPretS (Srpski PRETraživački Sistem)) is depicted in Figure 1.

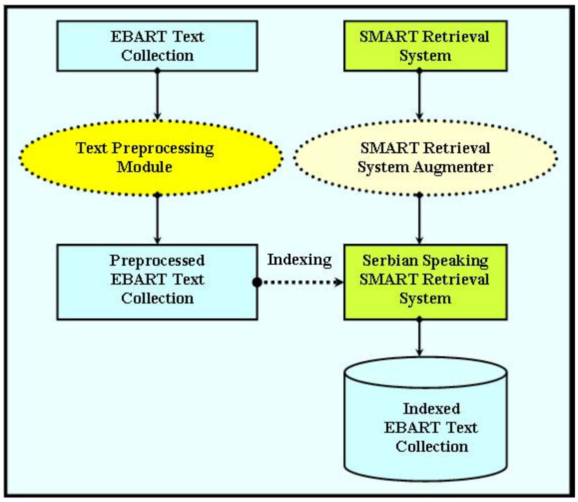
EBART Text Collection
SMART Retrieval System
Text Preprocessing Module
SMART Retrieval System Augmenter
Preprocessed EBART Text Collection
Indexing
Serbian Speaking SMART Retrieval System
Indexed EBART Text Collection

Figure 1: SPRETS' Overall Structure

Both, the text collection and SMART system are modified and preprocessed in order to make a proper indexing process possible. The preprocessed text collection is then fed into augmented SMART retrieval system enabled for Serbian language processing. Indexing that follows produces an index of the EBART text collection. Preprocessing of the text collection comprises of three main phases: (i) conversion of non-ASCII Serbian letters (*ć*, *č*, *đ*, *š*, and *ž*) into their corresponding ASCII transcripts (*cx*, *cy*, *dx*, *sx*, and *zx*), (ii) stopping, and (iii) word conflation.

While conversion of non-ASCII letters and stopping are rather straightforward operations, design and development of a successful word conflation algorithm is not as much. In addition, it has considerably reaching consequences as far as functioning of the retrieval system is concerned.

Serbian language possesses a rather complex morphology: seven noun cases, three noun genders, three noun numbers, six verb tenses, six verb forms, etc., etc. ([23], [46]). It was quite reasonable to assume that the quality of word conflation would have essential implications on retrieval process. This was eventually proved by our subsequent investigation, as well as the assessment and evaluation of the system's performance.

Thus, in addition to development of our retrieval system, we decided to experiment with it and measure the impact of different conflation modules. Thus, we set up a threefold experiment to run the system and measure its performance first without a conflation module, then using a stemmer with an exhaustive set of rules, and finally a stemmer with only a rudimentary stemming tenets. Consequently, we developed two algorithms for conflation modeled on some well known predecessors in the field ([1],[3], [9], [36], [39], [40], [41], [51]).

The Exhaustive Conflation Algorithm (ECA) conflation algorithm consists of about fifty rules of transformation that include stemming of word endings but also prefix analysis and letter substitutions in the middle of the word. A typical rule consists of its label, word measure condition (very short words are not transformed), word form description followed by a transformation (format established for STELEMMIN, generic minimal stemmer/lemmatizer ([36])). Since transformation can be done anywhere in a word, position of transformation is needed as well as descriptions of pattern to be replaced and pattern to replace it. All descriptions are regular expressions. An example of a typical word transformation rule would be as follows:

31; M>3; word has a at the next to last position preceded by a consonant and not followed by ć, č, đ, š, ž, lj, nj, or dž; remove the a; replace it with ''.

This method attempts to recognize word similarities not only when words differ in gender, number and tense but also when complex sound alterations take place. In Serbian, those include various linguistic phenomena as palatalization, *l* into *o* conversion, disappearing *a*, etc ([23], [46]).

Palatalization rule for instance implements conversion of *k*, *g*, *h* sounds into *č*, *ž*, *š* when they are found in front of vowels (e.g. nominative case *dru**g*** transforming into vocative *dru**ž**e*).

Rule implementing conversion of *l* into *o* is used when gender of adjectives or nouns changes from feminine to masculine or neutral (e.g. nominative feminine *ce**l**a* and nominative masculine *ce**o***).

The *disappearing a* rule deletes next to last *a* in a word if *a* is preceded by a consonant and is not followed by *ć*, *č*, *đ*, *š*, *ž*, *lj*, *nj*, or *dž* (e.g. genitive singular *manjka* derived from nominative singular of the noun *manj**ak***).

In another example, word *matematički* gets transformed into *matematik* by first applying the rule that removes ending *ki* and then another rule that replaces *č* into *k*.

Word *najjači* is transformed into *jk* by applying the following rules in order: deletion of prefix *naj*, deletion of a vowel at the end of a word, replacement of *č* by *k*, deletion of *a* at the next to last position when surrounded by consonants.

Because of the exhaustive nature of the algorithm, it succeeds in transforming a great majority of words into their corresponding common stems. However and expectedly, cases of excessive normalization were observed, as well. For example, word *pruge* (meaning *railways*) and *prugasti* (meaning *stripy*) get transformed into the same word *pruga* (*railway*). And, because of alike cases of excessive normalization, we decided to experiment with an alternative algorithm. Leaving out rules observed to be responsible for the before mentioned cases lead us to our next algorithm.

The Rudimentary Conflation Algorithm (RCA) was developed by reduction of the previous one to only a small subset of its original rules. Here, rules that deal with previously brought up complex linguistic phenomena are left out together with some other rules that were recognized to cause excessive stemming. When processing query *pruge Srbije* (*railways of Serbia*), ECA conflated words *pruge* (*railways*) and *prugaste* (*stripy*) into the same root word and consequently produced rather amusing but serious retrieval errors: it returned an article that dealt with popularity of Barbies in *stripy* bathing suites. In another similar case, given query *kirija* (*rent*), EC algorithm reduced words *kirija*, *Kira* (a first name) and *Kiri* (last name) to a same stem.

New RC algorithm is free of excessive normalization and capable to distinguish between and produce different stemmed forms in cases analogous to our previous examples (*pruge* and *prugaste*; *kirija*, *Kira* and *Kiri*). It overall showed a greater precision as detailed below.

With respect to the search without any word conflation preprocessing (our 'ground zero' procedure), general assessment is that both algorithms showed a fundamental improvement in both recall and precision. While the exhaustive conflation algorithm (ECA) exhibited an essential increase in recall and a good quality growth in precision, the rudimentary conflation algorithm (RCA) revealed a decent rise in recall and a superior increase in precision.

As we mentioned earlier, a two gigabyte EBART assortment of Serbian newspaper articles from years 2003-2007 was used as our document collection. We made an effort to follow the general recommendations of the TREC guidelines for IR systems' performance assessment ("http://www-nlpir.nist.gov/projects/tv2007/tv2007.html"). An evaluation experiment was set up to measure non-interpolated R-precision at EBART document counts (of 5, 10, 15, 20, etc.). Testing was done on 106 different queries compiled from a random sample of users. Precision was recorded for each individual query on each of the 'mod 5' document counts. As a final step, average individual query precisions at each of the document counts were calculated. Table 1 shows the document count precision averages along with an overall average per all document counts and percent increases of EC and RC algorithms with respect the base 'ground zero' search.

Document Count	Ground Zero Precision	ECA Average Precision	RCA Average Precision
5	0.698113	0.745283	0.822642
10	0.600943	0.676415	0.754717
15	0.520161	0.632704	0.701258
20	0.478338	0.575977	0.643904
...	...	...	...
Query Average	0.57439	0.65759	0.73063
% Increase	–	14.5 %	27.2 %

Table 1. R-precision statistic

(2) "POS Tagging and Shallow Parsing in a Less Resourced Language: A Case Study of Serbian", by Miroslav Martinović, Mladen Nikolić (submitted, and presently considered for CICLING 2008 conference to be held in Haifa, in February 2008).

This paper presents an experiment of building a POS tagger and a shallow parser in a language modest in language technology resources like Serbian. Approach designed and adopted is generic, flexible and driven by external configurations where phrases are described using a regular expression like formalism. As its backbone, TnT tagger was used. Since the corpora used for training lacked gender and number information only POS markups was utilized when building shallow parser. Annotated EBART collection of Serbian newspaper articles was used as corpus. Testing our tools demonstrates that their performance was comparable to equivalent tools in main stream well-resourced languages.

Part-of-speech tagging (POS tagging, POST or grammatical tagging) is the process of annotating words in a text as corresponding to a particular part of speech. Selected tags are based on both word's definition, as well as on its context (relationship with adjacent and related words in a phrase, sentence, or paragraph). POS tagging is nowadays almost exclusively done in the context of computational linguistics. The algorithms used associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags.

Shallow parsing (also known as chunking or "light parsing") is an analysis of a text which identifies the constituents (noun phrases, verb phrases, propositional phrases, etc.). However, this process is not expected to specify neither internal structure, nor the role of the phrases in the main sentence. It is a procedure extensively used in natural language processing.

Design and development of SPaS (*S*erbian *Pa*rt of *S*peech) tagger started by acquiring a high quality POS annotated Serbian corpora. This set in motion a sequence of actions driven by our choice to use the language independent TnT tagger. First step was a simple conversion of the EBART corpora into the format acceptable to TnT model generator. Following that, tagger's parameter generator is trained on the preprocessed annotated corpora. This in turn results in creation of model parameters by producing two new resources: Serbian corpus' lexical and contextual frequencies. With this, an end of the compilation phase has been reached. The two manufactured resources together with the TnT's tagger component embody an actual backbone resource of SPaS tagging system. During the run time, a new corpus (yet to get POS tagged) needs initially to get preprocessed into the format expected by the TnT system. It is then after fed into its component for tagging in conjunction with lexical and contextual frequencies produced during the compilation phase. The final output is POS tagged text.

Design of the SPaSer (*S*hallow *Pa*rser for *Ser*bian) was developed as a generic and a highly flexible language processing tool. It is driven by a configuration setting fed to it from without and containing definitions for all parseable phrases. The phrases are defined using regular expressions over parts of speech. Phrase definitions can also include previously introduced phrases (indicated by surrounding the ones used in definitions by '#' symbols).

Content of the configuration file used to define the simplest noun phrases (‘if’– from Serbian ‘*i*menska *f*raza), propositional phrases (‘pf’ – from Serbian ‘*p*redloška *f*raza), and verb phrases (‘gf’ – from Serbian ‘*g*lagolska *f*raza) in SPaSer:

```
<vrste_reci>
<vrsta id="1">imenica</vrsta>
<vrsta id="2">pridev< /vrsta>
<vrsta id="3">broj</vrsta>
<vrsta id="4">zamenica</vrsta>
<vrsta id="5">glagol</vrsta>
<vrsta id="6">prilog</vrsta>
<vrsta id="7">predlog</vrsta>
<vrsta id="8">veznik</vrsta>
<vrsta id="9">uzvik</vrsta>
<vrsta id="0">recca</vrsta>
<vrsta id=".">tacka</vrsta>
<vrsta id=",">zarez</vrsta>
<vrsta id="!">znak_uzvika</vrsta>
<vrsta id="?">znak_pitanja</vrsta>
<vrsta id="(">otvorena_zagrada</vrsta>
<vrsta id=")">zatvorena_zagrada</vrsta>
<vrsta id="nov_red">nov_red</vrsta>
<vrsta id="default">ostalo</vrsta>
</vrste_reci>
<fraze>
<fraza id="if" komentar="Imenske fraze">
        (prilog*pridev+)*imenica+
    </fraza>
    <fraza id="pf" komentar="Predloske fraze">
        predlog #if#
    </fraza>
    <fraza id="gf" komentar="Glagolske fraze">
        glagol #if#
    </fraza>
</fraze>
```

SPaSer expects an input collection together with an XML configuration file. As mentioned previously, configuration file uses regular expressions to define parseable phrases. Input text is subsequently POS tagged by SPaS tagger’s run-time component that actually does the annotation. Once the POS information is in place for the input text, configuration expressions are interpreted to recognize presence of parseable phrases in the text. Ultimately, the output consists of shallow parsed text denoted in XML.

Fragment of the SPaSer parsed text:

```
<if>
<imenica>godini</imenica>
</if>
<prilog>znatno</prilog>
<gf>
<glagol>popraviti</glagol>
<zamenica>njihov</zamenica>
<if>
<imenica>standard</imenica>
</if>
</gf>
<zarez>,</zarez>
<gf>
<glagol>pokazuje</glagol>
<if>
<imenica>istra`ivanje</imenica>
<imenica>agencije</imenica>
<imenica>Pressing</imenica>
</if>
</gf>
```

We evaluated SPaS tagger by measuring its accuracy (proportion of correctly tagged words among all tagged words). Annotated corpus was initially divided into a subset for training and another sub corpora used for testing (after all tags were stripped off). Correctness of SPaS’ tags was verified against the corresponding human annotated tags from the original testing set. The evaluation was done at different proportions for training data and is summarized in Table 1 below.

Training Data %	0.001%	0.01%	0.1%	1%	10%	50%	90%
Test Data Accuracy	53.6%	71.8%	85.8%	93.1%	97.1%	98.5%	98.9%
% of Known Words	26.9%	40.7%	59.5%	79.8%	92.6%	96.7%	97.7%
Accuracy on Known Words	99.6%	99.7%	99.4%	99.4%	99.5%	99.5%	99.5%
Accuracy on Unkn. Words	36.7%	52.6%	65.7%	68.2%	66.6%	69.8%	71.9%

Table 1. Evaluation Table for SPaS Tagger

The table includes accuracy for the testing set, as well as percentage of words in test corpus already known from the training corpus, with precisions on both, known and unknown words. Total number of words in the entire corpus was around 11,000,000.

Performance of the shallow parser was evaluated by measuring its precision and recall levels. Precision was assessed using two approaches. The first one (depicted best as ’binary’), scores 1 for each identified maximal phrase and then computes score percentage with respect to the entire set of phrases. An alternative scoring method (described best as ‘fuzzy’) accounts for any number of words from a phrase that the system identifies. For example, if a phrase consisting of two words has been identified, but the maximal phrase that includes it contains five words, the assigned score is 0.4 (or 2/5). Both approaches penalize presense of additional words that don’t belong to the phrase by assigning the score of 0.0.

Recall is determined analogously, the only difference being that instead of considering all phrases identified by the system, phrases that were supposed to be identified are counted. Text that was used in testing contained 142 noun phrases.

	Binary Scoring	Fuzzy Scoring
Precision	0.79	0.87
Recall	0.82	0.88

Table 2. Recall / Precision Evaluation for Shallow Parser

Our shallow parser is clearly a very legitimate tool. In cases when finding a phrase rather than finding the maximal variant of the phrase is a goal, SpaSer can be safely characterized as very successfull.

In addition, we noticed an interesting fact: there was almost no difference in performance when the parser was processing average size phrases (five or six words long) or very long phrases (with nine or ten words).

(3) “Vectorizing Structured Text with Fuzzy Evaluation for Topic Tracking in a Less Resourced Language: A Case Study of Serbian”, by Miroslav Martinović, Dušan Vasić (submitted, and presently considered for CICLING 2008 conference to be held in Haifa, in February 2008).

We address issue of modeling topic tracking in a language like Serbia. Adopted approach stems from and improves on classical vector model. We observed that web newspaper articles customarily exhibit a certain complexity of structure (retrievable date, author, title, frequently a subtitle and a supratitle, and even a general topic like Politics, Sport, Culture are commonly and directly obtainable). We tuned SMART retrieval system to take advantage of the assumed minimal text structure through a technique of structured vectorizing. A learning procedure acquired optimal weights for different article segments. Testing and evaluation were based on a notion that the cosine function of the Vector Model can be used to directly define a fuzzy tolerance relation. Cost from classical detection theory was then abstracted to a fuzzy level and allowed to specialize for relation costs. Corpus used in testing and evaluation consisted of web accessible articles from Serbian daily Politika.

System and algorithm featured in this paper, we chose to call SeToTraS (*Se*rbian *To*pic *Tra*cking *S*ystem). Fully implemented in Java, the system also includes packages developed as an object-oriented support for dealing with issues of vector modeled topic tracking in general. Development and implementation of this system followed a study and assessment on applicability of vector modeled systems to topic detection and tracking in the context of Serbian www texts. Vector space modeled (VSM) approach ([2], [5], [44]) was embraced as a starting point due to its simplicity, efficiency, and language independency, as well as because of the aptness it demonstrated while building SPretS, an information retrieval system for Serbian language.

Identifying shortcomings of the main stream vector model in this framework prompted development of a new algorithm which established promising improvements.

SeToTraS is a first, novel and unique full blown and complete TT system for Serbian language.

We have explored the concept of Topic Tracking by adopting VSM as the backbone of our system, as well as fuzzy logic for the evaluation and assessment.

We succeeded in insulating corpora gathering, corpora organization, term processing, text indication, vector normalization, annotation tables, algorithm testing/evaluation and machine learning. All major decisions were thus separated, to facilitate any future modifications, optimizations or even revolutionary twists in relations of project elements. For example, our algorithm for corpora gathering downloads specific web pages and extracts article structures and data therefrom, but is replaceable by any other analogous algorithm (e.g, SQL based, had we a database of articles).

We applied only the most basic algorithms for stemming and vector normalization. They both invite future refinements, with stemming being straightforwardly replaceable by lemmatization. Common experiences with SMART-like IR systems may also help to optimize the later element.

(VI) Acknowledgement

As part of a Sabbatical research investigating possibilities and limitations of transfer of NLP technologies from a resource rich to a less resourced language, the work on this paper has been funded in part by a grant awarded by Studenica foundation, as well as another grant by World University Service.

We would also like to thank the College of New Jersey for providing both logistical and financial opportunity for this research, as well as to Faculty of Mathematics at Belgrade University for being a gracious host of these activities. In particular, I'd like to express my gratitude to Dr. Vitas and Dr. Krstev for their willingness to help and advise.

Bibliography

1. Abney, S.: Part-of-Speech Tagging and Partial Parsing. In: Church, K., Young, S., Bloothooft, G. (eds): Corpus-Based Methods in Language and Speech (1996)
2. Allan, J.: Modeling Topics for Detection and Tracking. In: Pattern Recognition in Speech and Language Processing, Chou, W., Juang, F. (eds.), CRC Press (2002) 349-372
3. Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V., Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, R., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, E., Xu, J., Zhai., C. Challenges in Information Retrieval and Language Modeling. SIGIR Forum, March 2003.
4. Alemu, A., Asker, L., Getachew, M. Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward, in Proceedings of TALN 2003 Workshop on Natural Language Processing of Minority Languages and Small Languages, June, 2003.
5. Ariki, Y., Takao, S.: Study on New Term Weighting Method and New Vector Space Model Based on Word Space in Spoken Document Retrieval. In: Proceedings of the International Conference on Recherche d'Informations Assistee par Ordinateur(RIAO'00) 4 (2000) 116-131
6. Beigbeder, M., Mercier, A.: An Information Retrieval Model Using the Fuzzy Proximity Degree of Term Occurrences. In: Proceedings of the 2005 ACM symposium on Applied computing (2005) 1018 – 1022
7. Brants, T.: TnT -- a Statistical Part-of-Speech Tagger. In: Proceedings of the 6th Applied NLP Conference, ANLP-2000, (2000)
8. Cameron, R.D.: REX: XML Shallow Parsing with Regular Expressions. In: CMPT TR 1998-17, School of Computing Science, Simon Fraser University (1998)
9. Chrupala, G.: Simple Data-Driven Context Sensitive Lemmatization. In: Proceedings of SEPLN 2006 (2006)
10. Cucerzan, S., Yarowsky, D.: Bootstrapping a Multilingual Part-of-Speech Tagger in One Person-Day. In: Proceeding of the 6th Conference on Natural Language Learning Vol. 20 (2002) 1-7
11. Echihabi, A., Oard, D.W., Marcu, D., Hermjakob, U.: Cross-Language Question Answering at the USC Information Sciences Institute. In Proceedings of CLEF 2003: Cross-Language Evaluation Forum. Workshop N°4 Vol. 3237 (2003) 514-522
12. Fagundes da Silva, C., Osório, F.S., Vieira, R. Evaluating the Use of Linguistic Information in the Pre-processing Phase of Text Mining, in Proceedings of TALN

2003 Workshop on NL Processing of Minority Languages and Small Languages, June, 2003.

13. Fellbaum, C., editor, WordNet, An Electronic Lexical Database. MIT Press, 1998.

14. Gasperin, C., Vieira, R., Goulart, R., Quaresma, P. Extracting XML Syntactic Chunks from Portuguese Corpora, in Proceedings of TALN 2003 Workshop on NL Processing of Minority Languages and Small Languages, June, 2003.

15. Hajic, J., Cmejrek, M., Dorr, B., Ding, Y., Eisner, J., Gildea, D., Koo, T., Parton, K., Penn, G., Radev, D., Rambow, O. Natural Language Generation in the Context of Machine Translation. Technical Report, Center for Language and Speech Processing, JHU, 2002.

16. Hauptmann, A., Scheytt, P., Wactlar, H., Kennedy, P.E.: Multi-Lingual Informedia: A Demonstration of Speech Recognition and Information Retrieval across Multiple Languages. In: Proceedings of the DARPA Workshop on Broadcast News Understanding Systems (1998)

17. http://projects.ldc.upenn.edu/LCTL/

18. http://www.darpa.mil/ipto/Programs/gale/index.htm

19. http://www.darpa.mil/ipto/Programs/transtac/index.htm

20. http://www.nemlar.org/

21. http://www.telri.ac.uk/

22. Hughes, B.: Towards a Web Search Service for Minority Language Communities. In: Proceedings of Open Road Conference (2006)

23. Ivić, P., Pešikan, M., Klajn, I., Brborić, B.: Srpski jezicki prirucnik, Beogradska knjiga, Beograd (2007)

24. Jesperson, O., Language, its Nature, Origin and Development, George Allen & Unwin, London, 1921.

25. Kinyon, A.: A Language-Independent Shallow Parser Compiler. In: Proceedings of 10th EACL Conference (2001) 322-329

26. Korenius, T., Laurikkala, J., Jarvelin, K. and Juhola, M. "Stemming and Lemmatization in the Clustering of Finnish Text Documents", Proceedings of the 13th ACM International Conference on Information and Knowledge Management, Session IR-7, pp. 625-633, 2004.

27. Kraaij, W. and Pohlmann, R., "Porter's Stemming Algorithm for Dutch", Noordman LGM and de Vroomen WAM, eds. Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie, Tilburg, pp. 167-180, 1994.

28. Kraaij, W. and Pohlmann, R. "Evaluation of a Dutch Stemming Algorithm" Rowley J, ed. The New Review of Document and Text Management, Vol. 1, Taylor Graham, London, pp. 25-43, 1995.

29. Lam, W., Chan, K., Radev, D., Saggion, H., Teufel, S. Context-based Generic Cross-lingual Retrieval of Documents and Automated Summaries. Journal of the American Society for Information Science and Technology 56(2), February 2005.

30. Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., Thomas, S.: Relevance Models for Topic Detection and Tracking. In: Proceedings of the Human Language Technology Conference (HLT) (2002) 104-110

31. Leroy, G., Chen, H., Martinez, J. D.: A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text. In: Journal of Biomedical Informatics 36 (2003) 145-158.

32. Leuski, A., Allan, J.: Improving Realism of Topic Tracking Evaluation. In: Proceedings of ACM Conference on Research and Development in Information Retrieval (2002) 89-96

33. Li, X., Roth, R.: Exploring Evidence for Shallow Parsing. In: Proceedings of the Annual Conference on Computational Natural Language Learning (2001)

34. Martinovic, M., "Integrating Statistical and Linguistic Approaches in Building Intelligent Question-Answering Systems", Proceedings of the SSGRR 2002 International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet, 2002.

35. Martinovic, M., Curley, A., Gaskins, J. AARLISS – an Algorithm for Anaphora Resolution in Long-distance Inter Sentential Scenarios, In Proceedings of the 8th International Conference on Speech, Text and Dialogue, September 2005.

36. Martinovic, M., and Rofrano, L. "SteLemMin – A Generic Minimal Stem Algorithm for Word Conflation and Lemmatization", Proceedings of Workshop on Computational Modeling of Lexical Acquisition, 2006.

37. Martinovic, M., Sampath, G., Wagner, R., Briening, S. A Multilevel Text Processing Model of Newsgroup Dynamics : Implementation and Results, In Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems, NLDB'2003, 168-175, June 2003.

38. Martinovic, M., Vesic, S., Rakic, G.: Building an Information Retrieval System for Serbian – Challenges and Solutions. In: Proceedings of the 8th Annual International Interspeech Conference (2007)

39. Miller, G. WordNet : A Lexical Database for English. In C ACM, 38(1):49-51, 1995.

40. Paice, C. D., "Another Stemmer", ACM SIGIR Forum, Vol. 24, Issue 3, pp. 56-61, 1990.

41. Porter, M. F., "An Algorithm for Suffix Stripping", Program Vol. 4, No. 3, pp. 130-137, 1980.

42. Radev, D.R., Brew, C. editors. Effective Tools and Methodologies for Teaching Natural Language Process-

ing and Computational Linguistics, Philadelphia, PA, 2002.

43. Ratnaparkhi, A.: A Maximum Entropy Model for Part-Of-Speech Tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (1996)

44. Salton, G. (ed): The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall (1971)

45. Schmechel, N.: On the Lattice-Isomorphism between Fuzzy Equivalence Relations and Fuzzy Partitions. In: Proceedings of ISMVL-95 (1995)

46. Stevanović, M., Savremeni srpskohrvatski jezik, I, II, Beograd, 1994.

47. Thiele, H.: On the Mutual Definability of Fuzzy Tolerance Relations and Fuzzy Tolerance Coverings. In: Proceedings of the 25th International Symposium on Multiple-Valued Logic (1995) 140

48. Thiele, H., Schmechel, N.: On the Mutual Definability of Fuzzy Equivalence Relations and Fuzzy Partitions. In: Proceedings of the International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium (1995)

49. Voorhees, E. M., “Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness”, Information Processing and Management, 36(5), pp. 697-716.

50. Wayne, C. L.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. In Proceedings of LREC2000 (2000).

51. Xu, J., Croft, W.B. (1998) Corpus-based Stemming using Co-occurrence of Word Variants. In ACM Transactions on Information Systems 16(1), pp. 61-81. 1998.