Overcoming Data Sparsity in Automatic Transcription of Dictated Medical Findings

Edvin Pakoci
AlfaNum Speech Technologies
Novi Sad, Serbia
edvin.pakoci@alfanum.co.rs

Darko Pekar
AlfaNum Speech Technologies
Novi Sad, Serbia
darko.pekar@alfanum.co.rs

Branislav Popović
[1]*University of Novi Sad*
Faculty of Technical Sciences
Novi Sad, Serbia
[2]*Alfa BK University*
Academy of Arts
Belgrade, Serbia
bpopovic@uns.ac.rs

Milan Sečujski
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
secujski@uns.ac.rs

Vlado Delić
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
vdelic@uns.ac.rs

***Abstract*—This paper presents a method for introducing class *n*-gram language models as a means for overcoming data sparsity in the training of an automatic speech recognition (ASR) system aimed at transcription of dictated medical findings composed predominantly in the Serbian language, including occasional phrases in Latin. The classes used by the model are defined with the specific aim of avoiding the need of identifying an appropriate orthographic expansion of each abbreviation, number or other non-orthographic element in a particular context. Generated language models are decoded in Kaldi using token passing, and lattices generated in this way are rescored using recurrent neural network language models (RNNLM). Although the proposed approach requires extensive effort for initial definition of classes based on existing text corpora of medical findings, it improves the quality of the model and increases the degree of automation in the processing of future training corpora. As such, the proposed method is particularly suitable for training on noisy data, full of misspellings and other errors, such as medical findings. The feasibility of the approach has been tested on a corpus of medical findings in the domain of radiology, where a perplexity score of 59.55 and word error rate of 1.4% have been achieved.**

Keywords—the Serbian language, language modeling, class-based LM, code switching

I. Introduction

The accuracy and robustness of modern automatic speech recognition (ASR) systems largely depend on their ability to address the issue of data sparsity [1]. This issue is particularly challenging in case of general models with large vocabularies, as well as models that are required to handle different types of proper nouns, as these may not appear with sufficient frequency even in extremely large training corpora. The problem is further aggravated in case of synthetic languages, where inflection or agglutination are used to express syntactic relationships, which results in an extremely high number of distinct surface forms [2]. A common means for representation of out-of-vocabulary (OOV) words, as well as words otherwise inadequately modeled due to missing contexts in the training corpus, is a class-based *n*-gram language model [3]. This paper proposes a novel use of class *n*-gram models specifically aimed at challenges related to the development of an ASR system for transcription of dictated medical findings in the Serbian language. The principal challenges to be overcome are related to the training data, which consist of actual medical findings, and thus (1) typically exhibit an extremely high degree of ambiguity and noise, and (2) may include code switching to Latin, which is still widely used as lingua franca in medicine.

The remainder of the paper is organised as follows. Section 2 provides an overview of previous work relevant for the proposed method. Section 3 briefly discusses the specific challenges of the problem, principally reflected in the character of training data in the domain of medical findings, and gives an overview of the ASR system into which the developed language model is to be incorporated. Section 4 presents the proposed method in detail. Section 5 analyses the performance of the proposed method on a corpus of medical findings in the domain of radiology. Finally, Section 6 outlines the conclusions as well as the directions for future research.

II. Previous Work

The role of the language model (LM), as one of the principal components of any ASR system, is to describe the vocabulary and syntax of the language or speech domain in question, providing the system with allowed word sequences in limited vocabulary and grammar-based environments. Furthermore, LM also helps the acoustic model (AM) to decide on the correct word sequence by assigning costs (or scores) to different word sequences and identifying the word sequence with the lowest cost (highest score). Since unlikely word sequences are thus eliminated from the list of recognition result options, a well-trained LM can even compensate for certain impairments in the quality of the audio signal which affect intelligibility at the level of phonetic segments. In fact, it has been shown that a well-trained LM can bring the performance of an ASR system close to human speech recognition [4].

Statistical language models based on *n*-grams, i.e., estimated probabilities of individual word sequences of length up to *n*, have been the standard basis for language modeling in ASR for decades [5]. Although they have shown to be very effective in a wide array of applications, they are severely limited in their ability to efficiently model longer contexts. The mechanism of *N*-best list rescoring has been proposed in order to alleviate this problem [6], and it has been shown to significantly improve the performance of ASR systems. With the advent of machine learning, language models based on recurrent neural networks (RNN) [7, 8] or long short-term memory networks (LSTM) have been gaining popularity as tools for *N*-best list rescoring [9]. More recently, an Encoder-Classifier Model (EC-Model) has been proposed, based on direct comparisons between pairs in *N*-best lists [10], as well

This research was supported by the Science Fund of the Republic of Serbia, #6524560, AI S-ADAPT, as well as by Technology Transfer Experiment Open Call 2020 under the DIH-HERO innovation action (EU Horizon 2020 programme, grant agreement #825003, project MEDICTA).

as L2RS, treating *N*-best list rescoring as a learning problem and using information captured by state-of-the art NLP models [11]. There have also been attempts at modeling long-range semantic relations by focusing on semantic consistency within hypotheses, relying on continuous semantic representation based on BERT [12]. Besides their limitations in modeling long-range dependencies, *n*-gram models are also susceptible to data sparsity, and in many cases the probability of an *n*-gram cannot be directly estimated based on its frequency in a training corpus of any reasonable size. This has led to the development of a number of smoothing techniques, aimed at improving the robustness of probability estimates [6], as well as the introduction of class-based *n*-gram models [3].

Class-based *n*-gram LMs represent a common means for improving the robustness of language modeling by providing adequate representation of words or sequences not existing or underrepresented in the training corpus. Experiments on different test sets have confirmed that class *n*-gram LMs are successful in reducing perplexities as well as word error rates (WER) in speech recognition tasks [3, 13]. Word classes can be defined according to morphology, semantics or any other aspect of the word, and they can also be inferred automatically from data [14]. Class-based *n*-gram LMs also offer a simple way to add new words into the LM without any re-training or adaptation, since the problem is reduced to assigning a class to each new word. While early solutions typically required the user to manually assign the class to a new word [15], there has also been effort to automate this procedure. Most notable proposed solutions are based on the frequency of co-occurrence between the words in the vocabulary and the OOV word in question [3], morphological properties of these words [16] or cosine-similarity between corresponding word vectors [17].

III. Training Data and Related Challenges

The proposed method has been developed with the aim of being used within a speaker-independent ASR system aimed at automatic transcription of dictated medical findings in Serbian language, with occasional code-switching to Latin. The system, shown in Fig. 1, employs an initial LM based on class *n*-grams as well as an RNN LM used for rescoring of lattices obtained by decoding generated language models in Kaldi using token passing. The system represents an extension of the Serbian ASR described in [18, 19], trained on 904 hours of speech data, of which 379 hours (1087 speakers) of Serbian and 525 hours (1742 speakers) of Croatian, a phonetically and lexically extremely similar language. Namely, it was shown that a lower WER is achieved when the corpora in both languages are included in acoustic modeling, and that the slight differences in phonetic realizations of some consonants can be ignored.

In the proposed system pronunciation and language modeling are performed jointly for Latin and Serbian, not using general-domain text corpora used for training in [18, 19] but using only the corpus of medical findings described in this paper. Joint pronunciation modeling was possible owing to the fact that, at least among Serbian-speaking medical practitioners, the phonological inventory of Latin is actually a subset of the phonological inventory of Serbian. However, an additional dictionary of Latin had to be created in order to cover the terminology used in medical findings, due to the fact that, in spite of highly regular grapheme-to-phoneme conversion rules, the stress cannot be uniquely determined from the surface form of a word. Owing to the fact that the Latin terminology in medical findings is restricted to nominal grammar and that virtually no verbs are present in the data, a dictionary containing 3,287 lexemes (with appropriate declensions) was sufficient to cover all types with frequency of at least 3.

A. Overview of training data

The data available for training the model consists of a union of text corpora from various fields of medicine, predominantly radiology reports, containing a total of 14.8 million words. The data contains both official radiology reports and internal radiology reports, used for communication between experts within the same healthcare institution. Internal reports are typically composed in a far less formal language, contain great quantities of formal and informal abbreviations ("*St. post ICV aa V Hemiparesis l. sin.*", "*CT subependim. Ca++*"), may also contain proper names referring to colleagues or to other healthcare institutions, and are frequently written with far less attention to orthography. As the intended use of the ASR system includes composing internal reports as well, focusing on "clean" reports only and ignoring sections of the training corpus with particularly challenging data is not an option. Most of these difficulties can be overcome through judicious use of class-based language models, although not without significant effort of medical experts, whose involvement is necessary for initial class definition as well as pronunciation of Latin in some specific cases where available dictionaries offered no help. More detailed information about the training

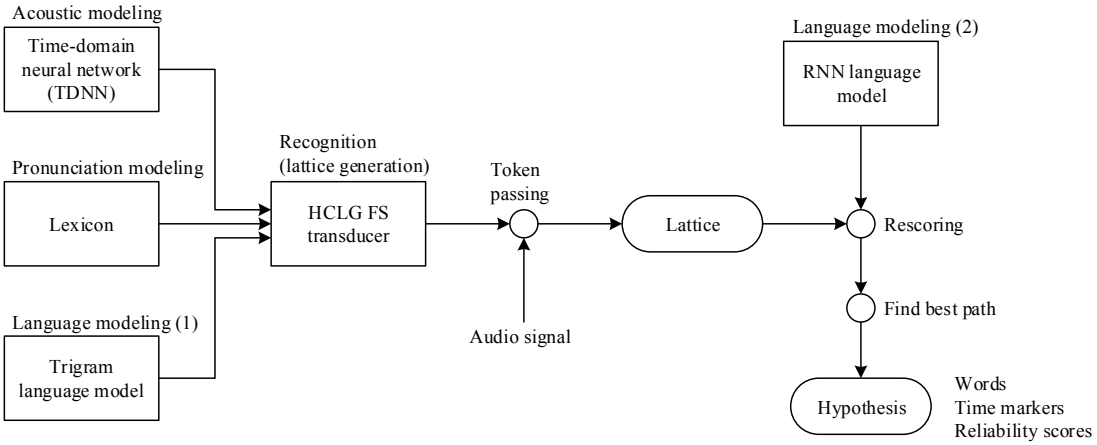

Fig. 1. Block diagram of the speech recognition system.

data is given in Section V, where it is compared with the versions after initial preprocessing and/or the proposed method for overcoming data sparsity has been carried out.

B. Class-based n-gram language models

In class-based language modeling the focus is on robust estimation of the probabilities of individual sequences of word classes rather than sequences of words. For that reason, the training corpus should contain either class sequences instead of word sequences, or a mixture of class and word sequences in case only some words have been assigned classes (e.g., proper nouns belonging to different semantic categories – first names, last names, names of places etc.) [20]. Assuming that the classes do not overlap, which is convenient as regards both the simplicity of the concept as well as its implementation, the standard word sequence probability for language modeling has to be modified. For example, the conventional *n*-gram formula for a word sequence $\mathbf{w} = \{w_1, w_2, \dots w_m\}$ is:

$$P(\mathbf{w}) = P(w_1) \prod_{i=2}^{m} P(w_i|w_{i-1}). \quad (1)$$

If an in-vocabulary word w_i is part of the class c_i (where w_i can be equivalent to c_i if it does not belong to any class), (1) needs to be rewritten into:

$$P(\mathbf{w}) = P(w_1|c_1)P(c_1) \prod_{i=2}^{m} P(w_i|c_i)P(c_i|c_{i-1}). \quad (2)$$

The within-class probabilities $P(w_i|c_i)$ can be estimated by dividing the number of occurrences of the word w_i in a given corpus with the total number of occurrences of all words within the class c_i in the same corpus. The paper proposes a novel use of class-based *n*-gram models, which enables efficient modeling of ambiguous abbreviations as well as infrequent words.

IV. Proposed Method

The principal problems addresed by the proposed method are (1) data sparsity resulting from an abundance of surface forms corresponding to a single semantic entity, as well as (2) the ambiguity of surface forms which may correspond to multiple semantic entities. For example, the Serbian adjective "suspektan" (eng. *suspected*) can have a multitude of surface forms, depending on its grammatical properties in a specific context. The same goes for the Latin adjective "suspectus" (with the same meaning), and both are frequently abbreviated into "*susp*". This leads to a situation which is ambiguous even as regards the language, and makes any attempt at expanding abbreviations into their orthographic form unfeasible. On the other hand, the ambiguity is further increased by the fact that, in this case, "*susp*" may stand for Serbian "suspenzija" (eng. *suspension*). More problematic examples can be easily provided, e.g. "IV" can stand for "intervertebral", "interventricular", "intraventricular", "intravenous" (in either Serbian or Latin, each with a multitude of different inflectional suffixes), as well as Roman numeral 4. The massive ambiguity related to the expansion of each non-orthographic token (abbreviation, number, etc.) into its orthographic form in a particular context is by far the most challenging problem related to the construction of language models based on text data of this type. We propose to overcome this ambiguity through the following procedure:

1) Text corpus is pre-processed by substituting abbreviated surface forms with class identifiers, whereby all identified potential surface forms corresponding to a class may be substituted even if they do not correspond to the same semantic entity. This leads to significant within-class ambiguity (e.g. all occurrences of any abbreviated form of "intraventricular" and "intravenous", as well as occurences of "IV" whether it stands for the Roman numeral or not, are assigned the same class). As will be explained in more detail later, classes can also span several words (e.g. abbreviated surface forms corresponding to either "arteria cerebri anterior" or "arteria communicans anterior" are assigned the same class, as both semantic entities are frequently denoted by the same abbreviation "ACA"). In other words, substitution is not necessarily performed at word level. It is clear that the initial definition of the content of each class requires extensive manual effort and the engagement of domain experts, but once defined, classes enable extremely fast processing of any new data once it is available.
2) A standard mixed class-based *n*-gram is built, and used in recognition. The most likely sequence of classes and non-classed words is identified, and the disambiguation between different realizations of each class is performed on the basis of the information obtained by the acoustic model as well as the relative frequency of each realization within the class. Appropriate text output is identified, based on the recognized semantic entity. A default output for each entity is defined in the class building phase (e.g. each instance of Latin "status" in the expression "status post" is abbreviated into "st."), and can be modified according to user preferences.

Several points of interest should be explained in more detail. Firstly, "all identified potential surface forms" in Step 1 refers not only to forms present in the corpus but also to all possible forms identified by the morphological analyzer [21], which is important from the point of view of future extension of training data. Namely, grammatical forms absent from the corpus in its current form may exist in new corpora. Secondly, the fact that classes can span several words prevents unnecessary introduction of ambiguity. For instance, the class *_intervertebral_disc* contains all identified surface forms of that semantic entity in either language, including many of those where "intervertebral" has been abbreviated into "i.v." or "IV". For that reason, all instances of "IV disc" in the training corpus containing abbreviations will be substituted with the class identifier *_intervertebral_disc* in Step 1. This will effectively reduce the ambiguity with respect to the case in which the instance of "IV" were to be substituted with the class *_IV*, also containing instances of "intravenous", "intraventricular" etc. Thirdly, semantic classes are used to handle different morphological forms of male first names, female first names, last names, cities, other settlements, names of medicines and chemical compounds. Finally, it should be noted that the assignment of words into classes may or may not include the assignment of all morphological forms of otherwise non-ambiguous words to a class, e.g. all inflected forms of the Serbian adjective "intervertebralni" (eng. *intervertebral*). Assignment to a class would be beneficial from the point of view of data sparsity, as all instances of an infrequent semantic item would be jointly represented, but it would basically discard morphological information carried in each individual realization and potentially increase the word error rate. For instance, the syntagm "intervertebral disc" in Serbian can have multiple surface forms, with the adjective and the noun agreeing in gender, case and number in each of them ("intervertebralnog diska", "intervertebralnom disku" etc.). However, if all

instances of the adjective were assigned the same class *_intervertebralni*, the ASR system, having recognized the class, would have to rely only on acoustics and word frequencies to decide which specific adjective form was used. As the suffix is not always easy to identify based on the acoustics only, this would lead to errors such as "intervertebralnog disku", which implies that, at least, class assignment should not be performed on the entire corpus, but that a sufficient number of actual surface forms should be retained. For that reason, substitution of multi-word expressions by class identifiers described in Step 1 is performed only in case at least one of the individual words is identified as out-of-vocabulary (i.e. not representing a full orthographic form).

V. RESULTS

A. Overview of the data

The initial corpus (hereafter referred to as C1) includes 14.8 million words, of which between 1.22% and 1.24% are Latin (the exact percentage is difficult to establish owing to occasional ambiguity at the language level). This corpus has first been preprocessed in order to eliminate automatically detectable orthographic errors (using available tools for Serbian) and perform basic unification principally regarding punctuation and measurement units. The obtained version of the corpus will hereafter be referred to as "unclassed" (C2). This version is then submitted to the classing procedure described in Section IV, and the "classed" version (C3) is obtained. An overview of the three versions of the corpus is given in Table I. It can be seen that preprocessing has drastically reduced the number of types, indicating the importance of error correction and unification. The number of both tokens and types has decreased as a result of classing, and it can also be noted that there are no abbreviations (lowercase OOV words followed by a period) left in C3. An overview of the content of the obtained semantic classes is given in Table II, but it should be noted that in case of frequent semantic entities such as medicines and chemical compounds, separate classes are actually used for each morphological form (e.g. *_medicine_nominative*, *_medicine_genitive* etc.), according to the dictionary [22]. Table III gives an overview of the frequency distribution in the training data. Apparent inconsistencies are due to various language specific details, most of which could not be elaborated here for brevity. For instance, a significant decrease in the number of acronyms in C3 is largely due to the fact that acronyms in Serbian frequently have suffixes indicating the grammatical case ("CT-a" vs. "CT-u", "CT-om"...), but these suffixes are lost in C3 since they do not usually appear in official medical findings. Besides a small number of semantic classes discussed above, C3 contains a total of 1261 classes, the largest one covering as many as 80 different surface forms which share the abbreviation "inf." ("infusion", "infection", "inferior", "inflammation", "infiltration"...).

B. Comparison of language models

The evaluation of the proposed research includes the calculation of the perplexity score of the model as well as the accuracy of ASR in an actual recognition task. However, it should be noted that in this research it has been impossible to establish a standard baseline such as the perplexity score of a non-class *n*-gram model or the word error rate (WER) of an ASR system based on such a model. Namely, as explained before, the proposed method avoids the need to expand each individual non-orthographic element in training data according to its context. For example, most words in "*St. post ICV aa V Hemiparesis l. sin.*" are never expanded to "*status post insultum cardiovascularem...*" in the training data, but they are substituted with class identifiers instead. Consequently, it is impossible to build a non-class *n*-gram language model or an ASR system based on it since there is no text corpus on which such a model could be trained. In these circumstances, evaluation is limited to establishing the achieved perplexity score of the proposed model as well as the accuracy of the ASR based on it, and comparing the obtained values with representative examples from literature. The results of this comparison can be interpreted only as general indicators, having in mind that datasets and various model settings may all be different.

TABLE I. OVERVIEW OF TRAINING DATA

Total number	Training corpus version		
	C1	C2	C3
Tokens	14.77 M	15.40 M	11.91 M
Types	290,106	112,751	92,987
Abbreviations 1[a]	760,411 (5.14 %)	291,722 (1,89%)	11,575 (0.10%)
Abbreviations 2[b]	60,915 (0.41%)	75,614 (0,49%)	0
Acronyms[c]	968,806 (6.56%)	720,208 (4.68%)	24,302 (0.20%)

[a] OOV tokens of up to 3 letters ("*tbl", "mg", "g", "Na"...*).
[b] OOV tokens, lowercase, followed by a period ("*caps.", "god.", "hosp."...*)
[c] OOV tokens, all caps (*"ECG", "PCR", "CT"...*)

TABLE II. OVERVIEW OF SEMANTIC CLASSES

Semantic entity	Tokens	Types
Male given name	717	128
Female given name	1599	107
Surname	5851	509
City	3432	34
Town/village	555	33
Medicine	163,877	776
Chemical compound	26,568	181

TABLE III. TYPE DISTRIBUTION IN EACH VERSION OF THE CORPUS (NUMBER OF TYPES THAT OCCUR AT LEAST *N* TIMES)

Frequency (*N*)	Training corpus version		
	C1	C2	C3
100,000	10/0/0[a]	20/0/0	17/0/0
10,000	282/0/0	236/0/0	176/0/0
1,000	1,682/108/87	1,240/59/77	998/2/3
100	7,826/327/416	5,637/260/256	4,724/9/22
10	32,745 /963/1738	1,8731/917/724	15,789/70/151

[a] Numbers indicate all words, abbreviations and acronyms respectively.

The perplexity score was measured on an independent dataset consisting of 10,000 sentences randomly drawn from C2 and withheld during training. A perplexity score of 59.55 was obtained on the full vocabulary, and in case the vocabulary was restricted to types occuring at least 6 times in the training data (a total of 20,960 types), it was reduced to 45.62. For the sake of comparison, we report perplexity scores obtained by existing methods for language modeling used in medical dictation. For instance, research described in [23] reports perplexity scores from 102 to 327 on the full vocabulary, obtained via within-domain interpolation with literal and semi-literal corpora, using 25 million words for training. On the other hand, [24] reports perplexity ranging from 185 to 613, but using only 270,000 words of training data. However,

as it is well known, the degree to which an ASR system benefits from a language model is not necessarily correlated with the decrease in perplexity [25]. Direct evaluation in a speech recognition task reveals the interaction between AM and LM, offering a much more practical insight. For that reason, we tested the accuracy of the ASR based on the proposed LM on a set of 1,000 utterances (13,423 words) remotely obtained from speakers who read aloud existing medical findings using their own consumer quality microphones in relatively silent environments of their rooms or offices. The text output of the system was compared to the original text, and a remarkably high accuracy of 98.6% at word level (a WER of 1.4%) was established. The specific design of the system requires the users to dictate punctuation marks as well, and these were also counted as words or multi-word expressions when calculating the WER. According to [26], the values of accuracy at word level reported by other studies are highly variable, ranging from 84.5% to 99%, but most studies report rates between 92% and 98%.

Illustrative examples of medical findings used in this evaluation as well as word errors made by the system can be found at alfanum.ftn.uns.ac.rs/medicta. A large percentage of errors corresponds to misrecognition of grammatical suffixes, leading to ungrammatical output such as "intervertebralnog disku" instead of "intervertebralnom disku". Such errors are caused by excessive classing, since the system is forced to distinguish between two acoustically very similar surface forms which belong to the same class and thus are equivalent from the point of view of LM. This suggests that the introduction of an explicit morphological model may bring a significant further reduction of WER.

VI. Conclusions and Future Work

The evaluation of the proposed method has shown it to be a feasible means for overcoming both data sparsity and many other problems related to the specific character of training data in the domain of medical findings. Most notably, the problem of expansion of tokens representing abbreviations or other non-orthographic elements such as numbers and measurement units is completely avoided. Although the execution of the proposed method requires significant manual effort from experts in the initial definition of classes, it greatly simplifies processing of future training data, eventually leading to its complete automation. It should also be noted that, although the feasibility of the approach has been demonstrated on a corpus including code switching between Serbian and Latin, the approach is essentially language independent and can be used on monolingual corpora as well. Owing to the use of semantic classes, many types of new items (e.g. proper names related to a particular medical institution) can be easily added to the language model without need for retraining.

Our further research on this topic will include the combination with the proposed model with a morphological language model, in order to improve the modelling of multi-word expressions without increasing the risk of ungrammatical text output. Finally, a great improvement in flexibility would be achieved by a resourceful introduction of an open-vocabulary language model capable of learning new words.

References

[1] B. Allison, D. Guthrie, and L. Guthrie, "Another look at the data sparsity problem," in Text, Speech and Dialogue 2006, LNCS, vol. 4188, P. Sojka, I. Kopeček and K. Pala, Eds. Heidelberg: Springer, 2006, pp. 327–334.

[2] E. Pakoci, B. Popović, and D. Pekar, "Using morphological data in language modeling for Serbian large vocabulary speech recognition," Comput. Intell. Neurosci., vol. 2019, pp 1–8, 2019.

[3] P. F. Brown, V. J. Dellapietra, P. V. de Souza, J. Lai, and R. Mercer, "Class-based N-gram models of natural language," Comput. Linguist., vol. 18, pp. 467–479, 1992.

[4] J. T. Goodman, "A bit of progress in language modeling: extended version," Tech. Rep. MSR-TR-2001-72, Microsoft Research, Redmond, WA, USA, 2001.

[5] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?," Proc. of IEEE, vol. 88, no. 8, pp. 1270–1278, 2000.

[6] D. Jurafsky, Speech & Language Processing, 3rd ed., Pearson Education India, 2000.

[7] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in: Proc. of the 11th Annual Conf. of ISCA, 2010.

[8] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, and S. Khudanpur, "Extensions of recurrent neural network language model," in: Proc. of ICASSP, pp. 5528–5531, 2011.

[9] H. Erdogan et al., "Multi-channel speech recognition: LSTMs all the way through," in: Proc. of CHiME-4 workshop, pp. 1–4, 2016.

[10] A. Ogawa, M. Delcroix, S. Karita, and T. Nakatani, "Rescoring *N*-best speech recognition list based on one-on-one hypothesis comparison using encoder-classifier model," in: Proc. of ICASSP 2018, pp. 6099–6103, 2018.

[11] Y. Song et al., "L2RS: A learning-to-rescore mechanism for automatic speech recognition", arXiv:1910.11496, 2019.

[12] D. Fohr and I. Illina, "BERT-based semantic model for rescoring *N*-best speech recognition list," in: Proc. of INTERSPEECH 2021, pp. 1867–1871, 2021.

[13] R. Kneser and H. Ney, "Improved clustering techniques for class-based statistical language modeling," in: Proc. of EUROSPEECH 1993, pp. 973–976, 1993.

[14] I. Bazzi and J. R. Glass, "A multi-class approach for modeling out-of-vocabulary words," in Proc. INTERSPEECH 2002, pp. 1613–1616, 2002.

[15] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system," in Proc. of ICASSP 1991, vol. 1, pp. 305– 308, 1991.

[16] A. Pražak, P. Ircing, and L. Müller, "Language model adaptation using different class-based models," in Proc. of SPECOM 2007, pp. 449–454, 2007.

[17] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Class-based *n*-gram language model for new words using out-of-vocabulary to in-vocabulary similarity," IEICE Trans. Inf. Syst., vol. E95-D, no. 9, pp. 2308–2317, 2012.

[18] E. Pakoci and B. Popović, "Recurrent neural networks and morphological features in language modeling for Serbian," in Proc. TELFOR 2021, pp. 186–193, 2021.

[19] E. Pakoci, "Influence of morphological features on language modeling with neural networks in speech recognition systems," Ph. D. dissertation, University of Novi Sad, Serbia, 2019.

[20] E. Pakoci and B. Popović, "Methods for using class based n-gram language models in the Kaldi toolkit," in Proc. SPECOM 2021, pp. 492–503, 2021.

[21] M. Sečujski and V. Delić, "A software tool for semi-automatic part-of-speech tagging and sentence accentuation in serbian language," in Proc. IS-LTC 2006, pp. 226-229, 2006.

[22] M. Sečujski, "Accentuation dictionary of Serbian intended for text-to-speech synthesis," in Proc. of Digital Image and Signal Processing Conf. DOGS 2002, pp. 17–20, 2002.

[23] G. Savova, M. Schonwetter, and S. Pakhomov, "Improving language model perplexity and recognition accuracy for medical dictations via within-domain interpolation with literal and semi-literal corpora", in Proc. INTERSPEECH, pp. 206–209, 2000.

[24] S. Pakhomov, M. Schonwetter, and J. Bachenko, "Generating training data for medical dictations," in Proc. NAACL, 2001.

[25] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," Speech Comm., vol. 38, no. 1-2, pp. 19–28, 2002.

[26] T.G. Poder, J. Fisette, and V. Déry, "Speech recognition for medical dictation: Overview in Quebec and systematic review," J. Med. Syst., vol. 42, no. 5, pp. 89–96, 2018.