METHODOLOGICAL PITFALLS OF AUTOMATIC SPEECH RECOGNITION

MILAN GNJATOVIĆ
Faculty of Technical Sciences, University of Novi Sad, milangnjatovic@uns.ac.rs

NEMANJA MAČEK
SECIT security consulting; Faculty of Engineering Management, Union University, Belgrade, macek.nemanja@gmail.com

ZLATOGOR MINCHEV
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, zlatogor@math.bas.bg

Abstract: *The broad surveillance potential of automatic speech recognition has been recognized across a range of application domains. Although considerable research effort has been devoted to the research question of achieving robust large-vocabulary continuous-speech recognition, it is still a fragile technology. In this paper, we discuss a methodological desideratum in this field. The currently dominant approaches to speech recognition relate to Bayesian statistical inference methods based on hidden Markov models and n-grams, or, ever-increasingly, neural networks. The common point of these approaches is that they are primarily corpus-driven and thus essentially agnostic of a broader interaction context, which represents a methodological limitation. Finally, this paper provides an overview of selected aspects of our recent research on natural language processing aimed at overcoming these methodological shortcomings.*

Keywords: *context-dependent speech recognition, focus tree, hybrid approach*

1. INTRODUCTION

The specification and design of automatic speech recognition systems involve different parameters of variation (including the vocabulary size, the speech fluency, the noise level, and the speaker-class characteristics), but the most important applications of this technology relate to large-vocabulary continuous-speech recognition. Such systems are intended to recognize spontaneous speech from previously unknown people, in realistic conditions [1]. However, although large research effort has already been invested in this field, the state-of-the-art speech recognition systems are still too restrictive, showing considerable error rates when applied in adverse conditions [2]. Most errors occur at the signal level, when the user's utterance was not correctly recognized although it was within the domain, scope and grammar of the recognition system [3]. These recognition errors are usually attributed to inadequate acoustic or language modelling. Complementary to such views, we discuss that a serious methodological limitation of the currently dominant approaches to speech recognition lies at the methodological level, as they do not account for a broader interaction context.

The paper is organized as follows. Section 2 provides a brief overview of the conceptualization of context in statistical pattern-matching approaches. Section 3 discusses on why language corpora alone do not suffice to produce reliable speech recognition systems in a general case. An overview of selected aspects of our recent research on a hybrid approach to automatic speech recognition is provided in Section 4. The paper ends with Section 5.

2. CONTEXT IN STATISTICAL PATTERN-MATCHING PARADIGM

The statistical approaches to speech recognition relate to Bayesian statistical inference methods based on hidden Markov models and n-grams. The recognition task is conceptualized as finding the most probable word sequence $\widehat{W}$ for an acoustic signal X, in a given search space L:

$$\widehat{W} = argmax_{W \in L} P(X|W)P(W).$$

Probability $P(X|W)$ is the observation likelihood derived from an underlying acoustic model. Words contained in a vocabulary are typically represented as hidden Markov models whose states express phone-like units. A widely accepted approach to large-vocabulary speech recognition is to apply context-dependent phone models, e.g., triphone hidden Markov models that represent phones in a particular left and right contexts.

Probability $P(W)$ is the prior probability derived from an underlying language model. It estimates the probability of a given word sequence $W = w_1 w_2 \dots w_n$ by using n-gram models (most often bigrams and trigrams, due to practical reasons):

$$P(w_1 w_2 \dots w_n) = \prod_{i=1}^{n} P(w_i | w_{i-1} \dots w_{i-N+1}).$$

Without going further into formal and practical details (for these, the reader may consult [1,4]), it is important to note that acoustic and language models include contextual information only at the phone and sentence levels, respectively. This very limited account of context is the reason why speech recognition systems based on this paradigm lack the robustness towards unexpected topic shifts, noisy conditions, etc. Therefore, the following questions may be raised here: Are these technical deficiencies due to inappropriate corpora used to train acoustic and language models, rather than to the methodological approach? Could a more comprehensive corpus provide enough data to address these technical deficiencies? The answer is negative, as discussed in the next section (cf. also [5]).

3. WHY LANGUAGE CORPORA ALONE DO NOT SUFFICE?

Language corpora have undoubtedly an important role in the design of speech recognition systems. However, their role tends to be overstated in research practices. Thus, Chomsky strongly criticizes the prevalent understanding "that the only real object is a corpus of data and that by automated analysis … one can derive everything that's relevant about the language" [6]. He describes it as "a novel concept of science that has emerged in the computational cognitive sciences and related areas of linguistics" [6]. With this concept, "an account of some phenomena is taken to be successful to the extent that it approximates unanalyzed data" [6]. Still, it has been widely adopted in recent language acquisition studies [7,8] and statistical approaches to machine learning.

Leaving this intellectual divide aside, there is a consensus that a language corpus should be representative, balanced, appropriately sized, etc. Nevertheless, these criteria are too vague and observer-relative. Although it is clear what representativeness of a corpus should mean, questions like "how do we identify the instances of language that are influential as models for the population?" still do not have definite answers [9]. In fact, we have no means to ensure or even evaluate the representativeness of a corpus [10, p. 57]. Similarly, a corpus is "pronounced balanced when the proportions of different kinds of text it contains should correspond with *informed* and *intuitive* judgments" [9] (emphasis added by author). A rigorous definition of these criteria is not to be expected soon.

In addition, speech recognition corpora usually contain recordings of utterances isolated from an interaction context, or telephone conversations (e.g., collected from a call centre, etc.) whose structure is objective-driven and thus not representative. In general case, the dialogue structure is not given beforehand, but evolves as the conversation unfolds [11,12]. Therefore, it is not even possible to produce a language corpus that would contain all relevant dialogue phenomena [13].

4. FROM RECOGNITION TO UNDER-STANDING

To overcoming the above methodological shortcomings, we apply a hybrid approach to speech recognition, incorporating both symbolic and statistical approaches. On the symbolic side, we refer to the focus tree model [14-18]. It is a symbolic and cognitively inspired model of attentional information in human-machine interaction that addresses the problem of robust recovery of semantic information from spontaneously uttered user's commands without explicit syntactic expectations. This model integrates three lines of research:

- the neurocognitive understanding of the focus of attention in working memory,
- the notion of attention related to the theory of discourse structure in the field of computational linguistics,
- the investigation of a corpus that comprises recordings of spontaneous speech-based human-machine interaction.

A corpus-based investigation of user commands resulted in the following findings:

- Propositional content is expressed by frequent insertion of chunks that explicitly relate to entities from the currently salient focus space. We refer to these parts as to focus stimuli.
- At the surface level, focus stimuli are non-recursive phrases. However, at the level of dialogue structure, they carry information about the attentional state.
- The order of focus stimuli within an utterance is flexible, while the word-order within them is rather fixed.
- Interaction participants often share a non-linguistic context. Therefore, speakers sometimes intentionally omit to explicitly utter information related to the attentional state because they believe it is already known by the interlocutor.
- In the strategy for recovering from non-understanding, speakers try to help the interlocutor by explicitly referring – using a constituent negation – to entities from the current interaction domain that should not be in the focus of attention.

In other words, attentional information is clearly signalled in the spontaneously produced user commands, and that it may be used to robustly recover semantic information without introducing explicit syntactic expectations [14,15]. In addition, the focus tree model includes cognitively-inspired and context-dependent evaluation of the retrieval cost and the integration difficulty of user's dialogue acts [18].

For the purpose of improving speech recognition performances, the focus tree model is used for post-processing of recognition hypotheses. The main idea of the proposed hybrid approach is that speech recognition hypotheses obtained by a standard statistically-based speech recognizer are further evaluated with respect to their retrieval cost, integration difficulty, and lexical matching. The system reduces the set of recognition hypotheses in an iterative manner, according to the following criteria [4]:

- The system first selects the recognition hypotheses with the minimum semantic integration difficulty in the given context. It there are more than one such a hypothesis, the second criterion is applied.
- From the recognition hypotheses selected in the previous step, the system selects hypotheses that

are most informative in the given context, i.e., provide the maximum retrieval cost. If there are again more than one such a hypothesis, the third criterion is applied.

- From the recognition hypotheses selected in the second step, the system selects hypotheses with the maximum lexical matching with respects to the system's vocabulary. If there more than one such a hypothesis, the last criterion is applied.
- From the recognition hypotheses selected in the third step, the system selects the recognition hypothesis that is assigned the highest probability by the standard, statistically-based speech recognition system.

For a detailed algorithm and a discussion on a prototype system, the reader may consult [4]. For the purpose of illustration, Table 1 summarizes the evaluation results, showing that the hybrid speech recognizer integrating both statistical and symbolic approaches outperforms the statistically-based recognizer.

Table 1: Evaluation results [4].

Parameter	Statistical approach	Hybrid approach
Number of sentences	980	980
Word error rate (%)	7.4	1.2
Sentence error rate (%)	38.6	3.5

5. CONCLUSION

Although considerable research effort has been devoted to the research question of achieving robust large-vocabulary continuous-speech recognition, it is still a fragile technology. This paper briefly discussed a methodological desideratum in this field. The common point of currently dominant approaches to automatic speech recognition is that they are primarily corpus-driven and thus essentially agnostic of a broader interaction context, which represents a methodological limitation. This paper provided an overview of selected aspects of our recent research on natural language processing aimed at overcoming the methodological shortcomings.

Acknowledgments

The presented study was sponsored by the Ministry of Education, Science and Technological Development of the Republic of Serbia (research grants III44008 and TR32035), and by the intergovernmental network EUREKA (research grant E!9944). The responsibility for the content of this article lies with the authors.

REFERENCES

[1] D. Jurafsky and J.H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", second edition, Prentice-Hall, 2009.

[2] C.H. Lee, "Fundamentals and Technical Challenges in Automatic Speech Recognition", In Proceedings of the XII International Conference "Speech and Computer" (SPECOM'2007); Moscow State Linguistic University, Moscow, Russia, pp. 25-44, 2007.

[3] D. Bohus and A.I. Rudnicky "Sorry, I Didn't Catch That! An Investigation of Non-Understanding Errors and Recovery Strategies", In: Dybkjær L., Minker W. (eds) Recent Trends in Discourse and Dialogue. Text, Speech and Language Technology, vol 39. Springer, Dordrecht, pp. 123-54, 2008.

[4] D. Mišković, M. Gnjatović, P. Štrbac, B. Trenkić, N. Jakovljević, V. Delić "Hybrid Methodological Approach to Context-Dependent Speech Recognition", International Journal of Advanced Robotic Systems, Vol. 14, No. 1, 2017.

[5] M. Gnjatović, "Do Language Corpora Suffice to Develop Conversational Agents?", Zbornik radova sa 9. konferencije Digitalna obrada govora i slike, DOGS 2012, ISBN 978-86-7892-439-2, Kovačica, Serbia, pp. 24–27, 2012.

[6] N. Chomsky, "Language and the Cognitive Science Revolution(s)", Carleton University, April 8, 2011, http://chomsky.info/talks/20110408.htm, 2011.

[7] M. Tomasello, "Constructing a Language: A Usage-Based Theory of Language Acquisition", Harvard University Press, 2003.

[8] J.R. Saffran, "Statistical Language Learning: Mechanisms and Constraints", Current Directions in Psychological Science 12(4), pp. 110-4, 2003.

[9] J. Sinclair, "Corpus and Text – Basic Principles", M. Wynne (ed.) "Developing Linguistic Corpora: a Guide to Good Practice", Oxford: Oxbow Books: pp. 1-16, http://ota.ahds.ac.uk/documents/creating/dlc/chapter2.htm, 2005.

[10] E. Tognini-Bonelli, "Corpus Linguistics at Work", John Benjamins, Amsterdam, 2001.

[11] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse", Comput. Linguist. 12(3), pp. 175-204, 1986.

[12] J. Searle, "Conversation", in: H. Parret, J. Verschueren (eds.) "(On) Searle on Conversation", John Banjamins Publishing Company, Philadelphia/Amsterdam, pp. 7-29, 1992.

[13] Y. Wilks, "Is There Progress on Talking Sensibly to Machines?", Science 318(5852), pp. 927-928, 2007.

[14] M. Gnjatović, M. Janev, and V. Delić, "Focus tree: Modeling attentional information in task-oriented human-machine interaction". Applied Intelligence 37(3), pp. 305-20, 2011.

[15] M. Gnjatović, V. Delić, "Cognitively-inspired representational approach to meaning in machine dialogue", Knowledge-Based Systems, Vol. 71, pp. 25-33, 2014.

[16] M. Gnjatović, B. Borovac, "Toward Conscious-Like Conversational Agents", In: Toward Robotic Socially Believable Behaving Systems, Volume II - Modeling Social Signals, A. Esposito, L.C. Jain (eds.), volume 106 of the series Intelligent Systems Reference Library, Springer, pp. 23-45, 2016.

[17] M. Gnjatović, "Changing Concepts of Machine Dialogue Management", in Proc. of the 5th IEEE

International Conference on Cognitive Infocommunications (CogInfoCom 2014), Vietri sul Mare, Italy, pp. 367-372, 2014.

[18] M. Gnjatović, “Therapist-Centered Design of a Robot's Dialogue Behavior”, *Cognitive Computation*, Vol. 6, No. 4, pp. 775-788, 2014.