Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 272–297

www.elsevier.com/locate/csl

Speech understanding for spoken dialogue systems: From corpus harvesting to grammar rule induction☆

Elias Iosif[a,b,*], Ioannis Klasinas[c], Georgia Athanasopoulou[c], Elisavet Palogiannidi[c], Spiros Georgiladakis[c], Katerina Louka[d], Alexandros Potamianos[a,b]

[a] *School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece*
[b] *"Athena" – Research and Innovation Center in Information, Communication and Knowledge Technologies, 15125 Athens, Greece*
[c] *School of Electronic and Computer Engineering, Technical University of Crete, 73100 Chania, Greece*
[d] *VoiceWeb S.A., 15124 Athens, Greece*

Received 9 September 2016; received in revised form 16 March 2017; accepted 15 August 2017
Available online 25 August 2017

Abstract

We investigate algorithms and tools for the semi-automatic authoring of grammars for spoken dialogue systems (SDS) proposing a framework that spans from corpora creation to grammar induction algorithms. A realistic human-in-the-loop approach is followed balancing automation and human intervention to optimize cost to performance ratio for grammar development. Web harvesting is the main approach investigated for eliciting spoken dialogue textual data, while crowdsourcing is also proposed as an alternative method. Several techniques are presented for constructing web queries and filtering the acquired corpora. We also investigate how the harvested corpora can be used for the automatic and semi-automatic (human-in-the-loop) induction of grammar rules. SDS grammar rules and induction algorithms are grouped into two types, namely, low- and high-level. Two families of algorithms are investigated for rule induction: one based on semantic similarity and distributional semantic models, and the other using more traditional statistical modeling approaches (e.g., slot-filling algorithms using Conditional Random Fields). Evaluation results are presented for two domains and languages. High-level induction precision scores up to 60% are obtained. Results advocate the portability of the proposed features and algorithms across languages and domains.

Keywords: Spoken dialogue systems; Grammar induction; Corpora creation; Semantic similarity; Web mining; Crowdsourcing

1. Introduction

Natural language understanding (NLU) is at the very heart of spoken dialogue systems (SDS) since its purpose is to transform the output of the speech recognizer into a semantic representation. Such representations are useful for other related tasks, e.g., the identification of speaker intention that drive the module of dialogue management. For example, consider an SDS for air tickets booking and the following example utterance: "I am leaving from Chicago".

☆ This paper has been recommended for acceptance by Roger Moore.
* School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece.
E-mail addresses: iosif.elias@gmail.com, iosife@central.ntua.gr (E. Iosif).

http://dx.doi.org/10.1016/j.csl.2017.08.002
0885-2308/

The salient part of this utterance is the lexical fragment "leaving from Chicago" that can be regarded as an instance of a grammar rule denoted as < DepartureCity > . Such grammar rules enable the understanding of the user input, e.g., the system can infer that 'Chicago is the departing city, and then proceed to other dialogue states for gathering any missing information, such as destination and travel dates. SDS grammars constitute a linguistic formalism that serves as the middleware between the recognized speech and the semantic representation. Speech understanding grammars can be distinguished into two broad categories, namely, finite-state-based (FSM) and statistical. Initial efforts in speech understanding grammar modeling were based on rule-based systems (e.g., Wang, 2001) suffering from poor generalizability and relying on manual updates (Pieraccini and Suendermann, 2012). Better results can be obtained using finite-state-based grammars (Potamianos and Kuo, 2000; Raymond et al., 2006), which enable the integration of automatic speech recognition output with NLU. More recent efforts rely on discriminative models such as Support Vector Machines (SVM) (Vapnik, 1998) and Conditional Random Fields (CRF) (Lafferty et al., 2001) and have been shown to outperform finite-state-based approaches (Raymond and Riccardi, 2007). Lately, top performance has been achieved by Recurrent Neural Networks (RNN) (Mesnil et al., 2015). The manual development of grammars poses an obstacle to the rapid porting of spoken dialogue systems to new domains and languages. The need for machine-assisted grammar induction has been an open research area for decades (Lari and Young, 1990; Chen, 1995) aiming to lower this barrier. Automatic (or semi-automatic) induction algorithms can be distinguished into two main categories, namely, resource-based and data-driven. The main drawback of resource-based approaches is the dependency on knowledge bases, which might not be available for under-resourced languages. This is tackled by the data-driven paradigm that relies (mostly) on corpora.

In this paper, we adopt a data-driven paradigm investigating various algorithms for the creation of text corpora and the induction of finite-state-based grammars. The end goal is to help automate the grammar development process. Unlike previous approaches (Wang and Acero, 2006; Cramer, 2007) that have focused on full automation, we adopt a human-in-the-loop approach where a developer bootstraps each grammar rule or request type with a few examples (seeds) and then machine learning algorithms are used to propose grammar rule enhancements to the developer. The enhancements are post-edited by the developer and new grammar rule suggestions are proposed by the system in an iterative fashion, until a grammar of sufficient quality is achieved. The main approach used for corpora creation is the harvesting of web data via the formulation of web search queries, followed by corpus filtering. The richness of the world wide web and its multilingual character enable the creation of corpora for less-resourced languages and domains. Note that the exploitation of web data is also appropriate for the development of statistical grammars where large amounts of data are required. In addition, various crowdsourcing tasks are used in order to elicit spoken dialogue text data. SDS grammar rules are distinguished into two types, namely, low- and high-level. Low-level rules refer to terminal concepts, e.g., the concept of city name can be represented as < City > → ("New York"|"Boston"). High-level rules are defined on top[1] of low-level rules, e.g., < DepartureCity > → ("fly from < City > "|"departing from < City > "). Two different families of language-agnostic induction algorithms are proposed, one for each type of rules. Greater focus is given to the induction of high-level rules, for which different approaches are proposed exploiting a rich set of features.

This work builds upon our prior research in Klasinas et al. (2013); Georgiladakis et al. (2014); Athanasopoulou et al. (2014); Palogiannidi et al. (2014), adding the following original contributions:

1. Regarding the harvesting of web data for corpora creation, two types of query generation (corpus- and grammar-based) are investigated, extending the work in Klasinas et al. (2013) where only the grammar-based approach was followed. In addition, here, more techniques for corpus filtering are proposed and compared. Detailed experimental results demonstrate that web harvesting is a viable approach for creating corpora intended for grammar induction.
2. In this work, we investigate the induction of both low- and high-level rules. Emphasis is given on the induction of high-level rules, a less researched area, unlike previous studies (Klasinas et al., 2013; Palogiannidi et al., 2014) that dealt only with low-level rules. We show that different similarity metrics and features are appropriate for the induction of low- and high-level rules. In total, four different approaches are proposed and compared for the high-level rule induction, extending the preliminary work in Athanasopoulou et al. (2014).

[1] High-level rules can be also stacked on top of each other, e.g., < DepartureArrivalCity > defined on top of < ArrivalCity > and < DepartureCity > .