Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 163–184

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

State of the art in statistical methods for language and speech processing☆

Jerome R. Bellegarda[a,*], Christof Monz[b]

[a] *Apple Inc., One Infinite Loop, Cupertino, CA 95014, USA*

[b] *Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

Received 14 November 2014; received in revised form 21 May 2015; accepted 3 July 2015

Available online 14 July 2015

Abstract

Recent years have seen rapid growth in the deployment of statistical methods for computational language and speech processing. The current popularity of such methods can be traced to the convergence of several factors, including the increasing amount of data now accessible, sustained advances in computing power and storage capabilities, and ongoing improvements in machine learning algorithms. The purpose of this contribution is to review the state of the art in both areas, point out the top trends in statistical modelling across a wide range of problems, and identify their most salient characteristics. The paper concludes with some prognostications regarding the likely impact on the field going forward.

Keywords: Statistical methods for language processing; Speech processing

1. Introduction

Statistical methods can be thought of as a way to leverage information primarily extracted from available raw data rather than derived from *a priori* expert knowledge. Over the past decade, such approaches have enjoyed rapidly increasing popularity in the field of computational language and speech processing. A wide spectrum of machine learning techniques have now been deployed to address the full complement of problems to be solved, from speech recognition and natural language modelling to information retrieval and text summarisation. This enthusiasm for statistical methods is the consequence of two main developments: (i) the explosion in the amount of data newly accessible, largely due to new social behaviours, societal transformations, as well as the vast spread of software systems and (ii) a steady reduction in the cost of computing and storage resources, which makes it possible to process increasingly large quantities of data with a reasonable amount of time and money.

☆ This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author. Tel.: +1 408 974 7647.

E-mail addresses: jerome@apple.com (J.R. Bellegarda), c.monz@uva.nl (C. Monz).

http://dx.doi.org/10.1016/j.csl.2015.07.001

0885-2308/

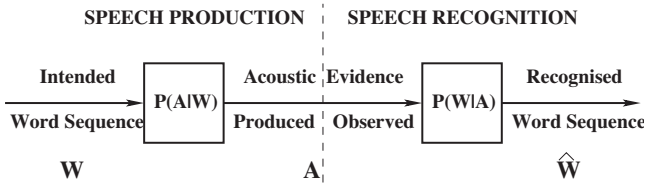

Fig. 1. Speech processing seen as information transmission over a noisy channel. The "Transmission Channel" is symbolised by the dashed line separating speech production from speech recognition.

1.1. Background

Arguably, the field of language and speech processing is inherently well suited for the kind of statistical methods now commonly adopted for data analytics under the "Big Data" paradigm. As early as the mid-1970s, speech processing began to be seen as an instance of information transmission over a noisy channel (Bahl et al., 1983). This view led to the well-known framework depicted in Fig. 1, where W refers to a word or sequence of words to be produced, A to the acoustic realisation of the resulting textual data, and $\hat{W}$ to the recognised sequence recovered from the observed evidence.

The framework of Fig. 1 comprises blocks labelled with parameters of the form $P(\cdot|\cdot)$, reflecting a noisy process characterised by a statistical model. The output sequence $\hat{W}$ then satisfies:

$$\hat{W} = \underset{W \in \mathcal{S}}{arg\max} P(W|A), \tag{1}$$

where the maximisation is done over all possible candidate word sequences in some feasible set $\mathcal{S}$. Using Bayes' rule and the fact that the maximisation is independent of the observation likelihood $P(A)$, (1) can also be written as:

$$\hat{W} = \underset{W \in \mathcal{S}}{arg\max} P(A|W)P(W), \tag{2}$$

which exposes two fundamental statistical models: the acoustic model $P(A|W)$, which characterises speech production, and the language model $P(W)$, which expresses the *a priori* probability of generating a particular sequence W in the language. These two models have formed the basis of automatic speech recognition (ASR) for the past three decades (Rabiner et al., 1996).

Interestingly, the same view can also encompass higher levels of language processing. For example, the framework of Fig. 2 corresponds to the "personal assistant" paradigm associated with products likeApple Siri (Apple Inc., 2011), Google Now (Google Inc., 2012), or Microsoft Cortana (Microsoft Corp., 2014). In that scenario, the user wants a particular problem solved, which is formulated as intent I. This intent is conveyed through a particular query sequence W, itself leading to an acoustic realisation as before. The recognised sequence $\hat{W}$ then elicits one of several possible actions, aimed at fulfilling the original intent I. Note that the blocks $P(W|I)$ and $P(I|W)$ play similar roles at the language level as the blocks $P(A|W)$ and $P(W|A)$ play at the speech level.

Given that both language and speech processing lend themselves particularly well to the application of statistical models, it is of interest to examine recent trends in each of the two research communities, along with their most salient characteristics.

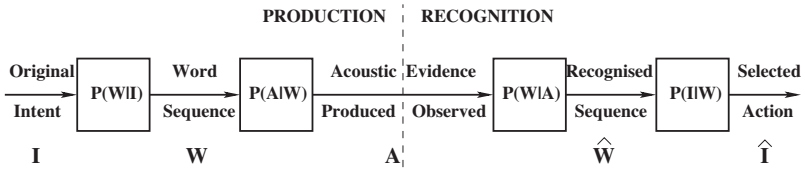

Fig. 2. An example of integrated language and speech processing: personal assistance seen as information transmission over a noisy channel.