Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 128 (2018) 55–64

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Natural Language and Speech Processing, ICNLSP 2015

Textual Data Selection for Language Modelling in the Scope of Automatic Speech Recognition

Freha Mezzoudj[a,b,*], David Langlois[d,e], Denis Jouvet[c,e], Abdelkader Benyettou[a]

[a] *Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, BP 1505, El M'Naouer, Oran, 31000, Algérie*
[b] *Université Hassiba Benbouali de Chlef, Ouled Fares, Chlef, 02000, Algérie*
[c] *MultiSpeech Group, LORIA, Inria, Villers-lès-Nancy, F-54600, France*
[d] *Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France*
[e] *CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France*

Abstract

The language model is an important module in many applications that produce natural language text, in particular speech recognition. Training of language models requires large amounts of textual data that matches with the target domain. Selection of target domain (or in-domain) data has been investigated in the past. For example [1] has proposed a criterion based on the difference of cross-entropy between models representing in-domain and non-domain-specific data. However evaluations were conducted using only two sources of data, one corresponding to the in-domain, and another one to generic data from which sentences are selected. In the scope of broadcast news and TV shows transcription systems, language models are built by interpolating several language models estimated from various data sources. This paper investigates the data selection process in this context of building interpolated language models for speech transcription. Results show that, in the selection process, the choice of the language models for representing in-domain and non-domain-specific data is critical. Moreover, it is better to apply the data selection only on some selected data sources. This way, the selection process leads to an improvement of 8.3 in terms of perplexity and 0.2% in terms of word-error rate on the French broadcast transcription task.

Keywords: Data selection; textual corpus; language model; cross-entropy; perplexity; speech recognition

1. Introduction

A Statistical Language Model constitutes one of the key components in several applications that produce natural language texts, such as large vocabulary speech recognition [2, 3], entity disambiguation[4], statistical machine translation[5], information retrieval[6], language text identification[7], handwriting recognition[8] and so on [9].

* Freha Mezzoudj. Tel.: +213 664-31-26-58 ; fax: +213 041-56-03-22.
E-mail address: freha.mezzoudj@univ-usto.dz

10.1016/j.procs.2018.03.008

The goal of Automatic Speech Recognition (ASR) is to accurately and efficiently convert a speech signal into a text message corresponding to the transcription of the spoken words, independently of the device used to record the speech, the speaker, or the environment [10]. The decoder, which is used to identify the pronounced words and sentences, exploits three types of knowledge corresponding to acoustic models which represent the acoustic realisation of sounds, to the lexicon which specifies the possible pronunciations of each word and to the language model (LM) which specifies the possible word sequences.

Our study focuses on French automatic speech transcription systems recently developed around projects and evaluation campaigns of automatic transcription of radio broadcasting programs. The initial ESTER1 campaigns of 2003 and 2005 [11] targeted radio broadcast news, the 2009 edition ESTER2 [12] introduced accented speech and news shows with spontaneous speech. The ETAPE 2011 evaluation [13] focused on TV material with various level of spontaneous speech and multiple speakers speech. The EPAC project of the French National Research agency (ANR) contributed to build the EPAC corpus of conversational speech manually and automatically transcribed [14].

The textual data that matches the best with this task are the textual data corresponding to the manual transcription of radio broadcast shows. This kind of textual data is costly to produce [15]. Hence, the amount of such training data is limited, and this impacts on the performance of a language model trained on this data only.

Besides the training data corresponding to the domain of interest, the development of language models can take benefit of data coming from other sources or other domains. When dealing with heterogeneous corpora, the conventional approach is first, to train an individual language model on each corpus (data source), and then, to combine them so as to maximize the fitting of the resulting (interpolated) language model with some development data representing the target task. Such an approach is used for our baseline LM, which is trained on a large text corpus of about two billion words. The text data comes from heterogeneous sources such as newspapers, news agency reports, web data, and a small quantity of manual transcriptions of radio broadcast programs.

Corpora may be noisy because of the variable quality of the sources. Noise data may lead to under-performing language models. To avoid such phenomenon, it might be useful to select a subset of relevant data for each source corpus.

This paper investigates methods for selecting data from textual corpora in view of improving language modelling for automatic speech transcription of broadcast news and TV shows. The selection methods used rely on computing, for each sentence, a score which represents how close the sentence is to in-domain data compared to non-domain-specific data. Several variants of scoring are proposed and analysed using perplexity measures. Finally, evaluations are conducted with respect to the automatic speech transcription of radio and TV shows.

The paper is organised as follows. Section 2 exposes the related work on data selection for language modelling and some associated techniques. Section 3 presents the experimental set-up, including the corpora used and the baseline LM. Textual data selection approaches and experiments are presented and discussed in Section 4. Section 5 presents conclusions and research directions for future work.

2. Data selection for language modelling

Classically, a high-performance language model is trained using a small corpus close to the target task (called in-domain data) and a huge data set not close to the task (called general domain data, or non-domain specific data). The in-domain data is carefully prepared by manual transcription. Unfortunately, this high-quality preparation leads to small data. Indeed, high-performance language models must deal with huge data in sake of coverage. To obtain huge data, one uses various sources easily available, often including web data. This leads to huge, but low-quality, general domain data.

This general domain data can contain relevant as well as irrelevant sentences with respect to in-domain data. The use of the irrelevant general domain data is probably more harmful than beneficial. In order to tackle this problem, various approaches were proposed and used in the literature to identify the most relevant portions of the general domain data prior to be used for training target LMs.

Klakow [16] uses a log-likelihood based criterion to select newspapers articles from a training corpus; and proposes two strategies for article removal. In the one-pass strategy, the criterion is computed for each article and then the top scoring articles are selected. The alternative is the iterative strategy, which, for each iteration, calculates the criterion for all articles and remove from the corpus a small fraction of worst scoring articles. This last strategy leads to a