Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 28 (2014) 1298–1316

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Paraphrastic language models☆

X. Liu*, M.J.F. Gales, P.C. Woodland

Cambridge University, Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England

Received 31 May 2013; received in revised form 28 March 2014; accepted 7 April 2014
Available online 30 April 2014

Abstract

Natural languages are known for their expressive richness. Many sentences can be used to represent the same underlying meaning. Only modelling the observed surface word sequence can result in poor context coverage and generalization, for example, when using *n*-gram language models (LMs). This paper proposes a novel form of language model, the paraphrastic LM, that addresses these issues. A phrase level paraphrase model statistically learned from standard text data with no semantic annotation is used to generate multiple paraphrase variants. LM probabilities are then estimated by maximizing their marginal probability. Multi-level language models estimated at both the word level and the phrase level are combined. An efficient weighted finite state transducer (WFST) based paraphrase generation approach is also presented. Significant error rate reductions of 0.5–0.6% absolute were obtained over the baseline *n*-gram LMs on two state-of-the-art recognition tasks for English conversational telephone speech and Mandarin Chinese broadcast speech using a paraphrastic multi-level LM modelling both word and phrase sequences. When it is further combined with word and phrase level feed-forward neural network LMs, a significant error rate reduction of 0.9% absolute (9% relative) and 0.5% absolute (5% relative) were obtained over the baseline *n*-gram and neural network LMs respectively.

Keywords: Language modelling; Paraphrase; Speech recognition

1. Introduction

Natural languages are known to have layered structures, a hidden and deeper structure that represents the meaning and core semantic relations within a sentence, and a surface form found in normal written texts or spoken language, as formulated in linguistic theories such as generative grammar Chomsky (1966), Jackendoff (1974). The mapping from the meaning to the observed surface form involves a natural language generation process. As multiple surface realizations can be used to convey identical or similar semantic information, this mapping is often one-to-many. These different surface realizations are paraphrastic to one another. They were created by using different syntactic, lexical and morphological rules in the generation process. Functionally these paraphrase variants represent different styles, dialects or other speaker specific characteristics. Due to their presence, only modelling the observed surface word sequence can result in poor context coverage, for example, when using standard *n*-gram language models (LMs).

☆ This paper has been recommended for acceptance by 'Riley Michael'.

* Corresponding author. Tel.: +44 1223 766512; fax: +44 1223 332662.
E-mail addresses: xl207@cam.ac.uk, xl207@eng.ac.uk (X. Liu), mjfg@eng.cam.ac.uk (M.J.F. Gales), pcw@eng.cam.ac.uk (P.C. Woodland).

http://dx.doi.org/10.1016/j.csl.2014.04.004
0885-2308/

One approach to handle this problem requires directly modelling paraphrase variants when constructing the LM. As alternative expressions of the same meaning are now considered, the resulting language model's context coverage and generalization performance is expected to improve. Along this line, the use of word level synonym features Cao et al. (2005), Hoberman and Rosenfeld (2002), Jelinek et al. (1990), Kneser and Peters (1997) has been investigated in early research for n-gram and class n-gram based Brown et al. (1992) language models. However, there are two issues associated with these existing approaches. First, the paraphrastic relationship between longer span syntactic structures, such as phrases, is largely ignored. A more general form of modelling that can also capture a higher level and longer span paraphrase mapping should be more effective. Second, previous research focused on using manually derived expert semantic labelling provided by resources such as WordNet Fellbaum (1998). As manual annotation is usually very expensive to produce, these methods cannot be applied to large corpora or languages without suitable WordNet-type resources. Hence, automatic, statistical paraphrase induction and extraction techniques are required.

In order to address these issues, this paper presents a novel form of language model, the paraphrastic language model (PLM). It provides a highly flexible and general form of paraphrase modelling that can be used at either the word, phrase or sentence level. The paraphrastic relationship between longer span syntactic structures can thus be effectively captured. A phrase level paraphrase model statistically learned from standard text data is used to generate multiple paraphrase variants for the training data. Language model probabilities are then estimated by maximizing the marginal probability of these variants. By linking language generation and modelling, paraphrastic LMs exploit an intuitive and interpretable parameter smoothing scheme to improve generalization performance. In order to leverage the complementary characteristics of paraphrastic LMs and feed-forward neural network LMs (NNLMs) Bengio et al. (2003), Kuo et al. (2012), Le et al. (2013), Park et al. (2010), Schwenk (2007), the combination between the two is also investigated.

This paper extends previous research summarized in Liu et al. (2012b, 2013c). A more complete study of using paraphrastic language models for speech recognition is presented. Various important aspects of this work, including the theory and implementation of the statistical paraphrase learning algorithm, the generation of paraphrase lattices and the construction of phrase and multi-level paraphrastic LMs, are covered in detail in this paper. These are further augmented by a full set of experimental results presented to demonstrate the advantages of paraphrastic LMs over existing modelling methods. This paper shows the applicability of paraphrastic LMs to multiple languages and genres, the scaling behaviour on varying amounts of training data, and their complementarity to other established language modelling techniques.

The rest of the paper is organized as follows. Paraphrastic language models are introduced in Section 2. A statistical n-gram phrase pair based paraphrase extraction scheme is presented in Section 3. Paraphrase lattice generation using a weighted finite state transducer (WFST) approach is described in Section 4. The estimation of paraphrastic LMs is presented in Section 5. The combination between paraphrastic LMs and feed-forward neural network LMs is proposed in Section 6. In Section 7 a range of paraphrastic LMs are evaluated on two state-of-the-art speech recognition tasks for English conversational telephone speech and Chinese broadcast speech respectively. Section 8 is the conclusion and possible future work.

2. Paraphrastic language models

As discussed above, in order to capture the paraphrase mapping between longer span syntactic structures, a more general form of modelling is required. To address this issue, the particular type of LMs proposed in this paper can flexibly model paraphrastic relationships at the word, phrase and sentence level. As LM probabilities are estimated in the paraphrased domain, they are referred to as *paraphrastic language models* (PLMs) in this paper. For a *surface word sequence* $\mathcal{W} =< w_1, w_2, \ldots, w_i, \ldots, w_L >$ of L words in the training data, for example, "*And I generally prefer*", rather than maximizing the surface word sequence log-probability $\ln P(\mathcal{W})$ as for conventional LMs, the marginal probability over its *paraphrase variant sequences*, $\{\mathcal{W}'\}$, such as "*And I just like*" or "*I mean I want*", is maximized Liu et al. (2012b, 2013c),

$$\mathcal{F}(\mathcal{W}) = \ln\left(\sum_{\boldsymbol{\psi},\boldsymbol{\psi}',\mathcal{W}'} P(\mathcal{W}|\boldsymbol{\psi})P(\boldsymbol{\psi}|\boldsymbol{\psi}')P(\boldsymbol{\psi}'|\mathcal{W}')P_{\mathsf{PLM}}(\mathcal{W}')\right) \tag{1}$$