Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 185–199

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

A unified framework for translation and understanding allowing discriminative joint decoding for multilingual speech semantic interpretation☆

Bassam Jabaian [a,*], Fabrice Lefèvre [a], Laurent Besacier [b]

[a] LIA-CERI, University of Avignon, Avignon, France
[b] LIG, University Joseph Fourier, Grenoble, France

Received 30 November 2013; received in revised form 5 June 2014; accepted 23 June 2014
Available online 2 July 2014

Abstract

Probabilistic approaches are now widespread in most natural language processing applications and selection of a particular approach usually depends on the task at hand. Targeting speech semantic interpretation in a multilingual context, this paper presents a comparison between the state-of-the-art methods used for machine translation and speech understanding. This comparison justifies our proposition of a unified framework for both tasks based on a discriminative approach. We demonstrate that this framework can be used to perform a joint translation-understanding decoding which allows to combine, in the same process, translation and semantic tagging scores of a sentence. A cascade of finite-state transducers is used to compose the translation and understanding hypothesis graphs (1-bests, word graphs or confusion networks). Not only this proposition is competitive with the state-of-the-art techniques but also its framework is even more attractive as it can be generalized to other components of human–machine vocal interfaces (e.g. speech recognizer) so as to allow a richer transmission of information between them.

Keywords: Multilingual speech understanding; Conditional random fields; Hypothesis graphs; Statistical machine translation; Dialogue systems

1. Introduction

Nowadays, probabilistic approaches are widely used in natural language processing (NLP) applications (speech recognition, machine translation, syntactic parsing, part-of-speech or semantic tagging). Many different approaches are available in the literature and their relative performances differ according to the targeted tasks. So, considering a specific task, it not always possible to know the best performing approach before evaluating most of the models. Though some important features of the various approaches may prevent us from testing all the combinations with respect to some of the task characteristics.

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +33 0490843534.
E-mail addresses: bassam.jabaian@univ-avignon.fr (B. Jabaian), fabrice.lefevre@univ-avignon.fr (F. Lefèvre), laurent.besacier@imag.fr (L. Besacier).

http://dx.doi.org/10.1016/j.csl.2014.06.003

For instance, for spoken language understanding (SLU), the task of extracting the meaning from a user's utterance, conditional random fields (CRF) (Lafferty et al., 2001) have been shown to be the most efficient approach so far (Hahn et al., 2010); whereas for statistical machine translation (SMT), the log-linear phrase-based statistical machine translation (LLPB-SMT) model (Koehn et al., 2003) is the most commonly used and has shown its potential for many language pairs and domains of application.

However despite the differences in their formal descriptions, distinctions between probabilistic approaches tend to fade away when confronted with practical considerations and the numerous assumptions required for their implementations. Some works already proposed the use of discriminative approaches, such as CRF, for SMT (Och and Ney, 2002; Liang et al., 2006; Lavergne et al., 2011) while at the same time the phrase-based translation pipeline was also investigated in the context of other natural language processing tasks such as grapheme–phoneme conversion (Rama et al., 2009) or part-of-speech tagging (Gascó i Mora and Sánchez Peiró, 2007).

So far, many works have considered the issue of multilingual systems for different NLP tasks such as cross-lingual information retrieval (e.g. Capstick et al., 2000; Jagarlamudi and Kumaran, 2008), cross-lingual information distillation (Singla and Hakkani-Tur, 2008), multilingual speech recognition adaptation (Schultz, 2004) and cross-lingual spoken language understanding (e.g. Minker, 1998; Lefèvre et al., 2010).

In this paper, our overall objective is to develop efficient approaches for speech understanding in a multilingual context (where SMT is also involved, as explained later). In this outlook, the state-of-the-art approaches are investigated for each of the underlying issues: CRF for speech understanding and LLPB-SMT for translation. In a first step, we propose to use and optimize the LLPB-SMT approach for speech understanding, and also to integrate a CRF-based model in a machine translation module to evaluate the practical interest of each method. This preliminary study serves to highlight the specificities of each task and to evaluate the performance of the respective approaches on these tasks.

Besides, we have shown in a previous work that the use of machine translation is an effective solution for the portability of an understanding system from a language to another (Jabaian et al., 2010). For one of the best performing configurations, the portability is simply obtained by cascading a translation module with an understanding system, the overall idea being to translate the user's inputs into a language for which we already have a performing SLU system. This idea is to be contrasted with this consisting in trying to build a performing system in the new language, for which we generally lack enough usable data.

But, in this context, it has been stated that the best understanding output is not always generated from the best translation hypothesis. From our experience, it is often due to bad word reordering. Therefore, the selection of the best translation is not sufficient to optimize the overall system in a multilingual understanding scenario. Consequently, based on the initial comparison made between both tasks, we propose a model that can jointly decode the user's inputs in terms of translation and understanding hypotheses. This joint decoding selects a translation taking into account the semantic tagging generated for this translation. It is no longer searched for the best possible translation, but for the translation that can be semantically labelled in the best possible way.

The reported experiments are carried out on the French MEDIA man–machine dialogue corpus (Bonneau-Maynard et al., 2006). Manually transcribed and conceptually annotated data allow to train an initial CRF-based understanding system for French. In order to use this system for Italian semantic tagging, an Italian to French translation system, based on either LLPB-SMT or CRF, is trained with few manual data. This system is used during decoding to infer translations (from Italian into French) in order to provide inputs to the French understanding system. We show how these models are merged in a single decoding loop by means of hypothesis graphs and combined scores. To define the performance upper-bounds of the proposed methods, all experiments are performed using a manual transcription of the speech data.

Preliminary results have already been presented in Jabaian et al. (2013b). This current version is not only a refinement of these experiments but it also provides a deeper comparison of the techniques at hand with more result analysis. For instance, we investigate more systematically the effect of the tuning mechanism used to optimize the performance of the joint models and show that all the results are noticeably improved. This optimization, based on the minimum error rate training (MERT) algorithm (Och, 2003), is first applied to each model (translation and understanding) separately, then it is generalized to perform an optimized joint decoding using both systems.

The paper is organized as follows: Section 2 presents the use of the log-linear phrase-based stochastic machine translation approach for speech understanding. Then Section 3 describes the use of Conditional Random Fields for machine translation and the parameter tuning of this model. Our proposal of a joint decoding process between translation