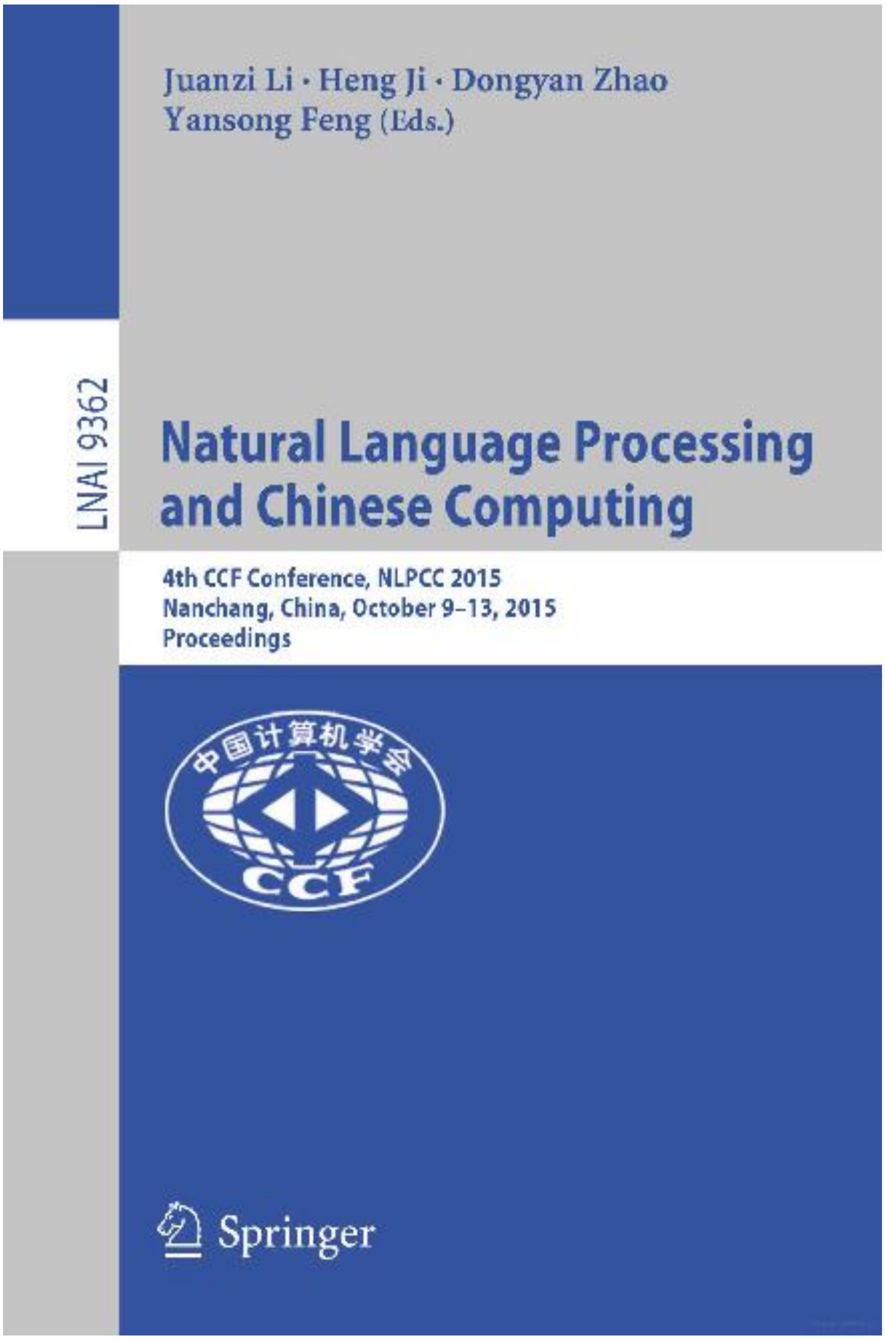

Juanzi Li · Heng Ji · Dongyan Zhao
Yansong Feng (Eds.)

LNAI 9362

Natural Language Processing and Chinese Computing

4th CCF Conference, NLPCC 2015
Nanchang, China, October 9–13, 2015
Proceedings

Springer

Editors

Juanzi Li
Tsinghua University
Beijing
China

Dongyan Zhao
Peking University
Beijing
China

Heng Ji
Rensselaer Polytechnic Institute
Troy, NY
USA

Yansong Feng
Peking University
Beijing
China

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-25206-3 ISBN 978-3-319-25207-0 (eBook)
DOI 10.1007/978-3-319-25207-0

Library of Congress Control Number: 2015951416

LNCS Sublibrary: SL7 – Artificial Intelligence

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Refining Kazakh Word Alignment Using Simulation Modeling Methods for Statistical Machine Translation

Amandyk Kartbayev(✉)

Laboratory of Intelligent Information Systems,
Al-Farabi Kazakh National University, Almaty, Kazakhstan
a.kartbayev@gmail.com

Abstract. Word alignment play an important role in the training of statistical machine translation systems. We present a technique to refine word alignments at phrase level after the collection of sentences from the Kazakh-English parallel corpora. The estimation technique extracts the phrase pairs from the word alignment and then incorporates them into the translation system for further steps. Although it is a pretty important step in training procedure, an word alignment process often has practical concerns with agglutinative languages. We consider an approach, which is a step towards an improved statistical translation model that incorporates morphological information and has better translation performance. Our goal is to present a statistical model of the morphology dependent procedure, which was evaluated over the Kazakh-English language pair and has obtained an improved BLEU score over state-of-the-art models.

Keywords: Word alignment · Optimization · Kazakh morphology · Word segmentation · Machine translation

1 Introduction

In this paper, we present the work done for improving a baseline statistical machine translation (SMT) system from an agglutinative Kazakh language to English. In this pair of translation, English word correspond to Kazakh suffixes that can fit more than one of the suffixes to the word. For instance, using the Kazakh lemma el - 'state' we can generate 'eldin' - 'of the state', 'elge' - 'to the state' and so on. The Kazakh language, which is the majority language in the Republic of Kazakhstan, has poor open resources and there are a small available parallel corpora unlike to other languages that more widely used as English or French. The parallel corpora for this work is 70k sentences for Kazakh and English, much bigger one than this corpus we had before.

In previous work[1], we described an approach to word alignment intended to address these problems. A research more relevant to that work was done by Bisazza and Federico[2]. The main goal of this research, different from the previous works, that is to make proposals that increase the expected benefit from

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 421–427, 2015.
DOI: 10.1007/978-3-319-25207-0_38