Van-Nam Huynh
Masahiro Inuiguchi
Thierry Denoeux (Eds.)

LNAI 9376

Integrated Uncertainty in Knowledge Modelling and Decision Making

4th International Symposium, IUKM 2015
Nha Trang, Vietnam, October 15–17, 2015
Proceedings

Springer

Van-Nam Huynh · Masahiro Inuiguchi
Thierry Denoeux (Eds.)

Integrated Uncertainty in Knowledge Modelling and Decision Making

4th International Symposium, IUKM 2015
Nha Trang, Vietnam, October 15–17, 2015
Proceedings

Springer

Editors
Van-Nam Huynh
Japan Advanced Institute of Science and Technology
Nomi
Japan

Masahiro Inuiguchi
Graduate School of Engineering Science
Osaka
Japan

Thierry Denoeux
Université de Technologie de Compiègne
Compiègne
France

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-25134-9 ISBN 978-3-319-25135-6 (eBook)
DOI 10.1007/978-3-319-25135-6

Library of Congress Control Number: 2015951823

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)

Learning Word Alignment Models for Kazakh-English Machine Translation

Amandyk Kartbayev(✉)

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

Abstract. In this paper, we address to the most essential challenges in the word alignment quality. Word alignment is a widely used phenomenon in the field of machine translation. However, a small research has been dedicated to the revealing of its discrete properties. This paper presents word segmentation, the probability distributions, and the statistical properties of word alignment in the transparent and a real life dataset. The result suggests that there is no single best method for alignment evaluation. For Kazakh-English pair we attempted to improve the phrase tables with the choice of alignment method, which need to be adapted to the requirements in the specific project. Experimental results show that the processed parallel data reduced word alignment error rate and achieved the highest BLEU improvement on the random parallel corpora.

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

1 Introduction

In recent years, the several studies were conducted to evaluate the relationships between word alignment and machine translation performance. The phrase table is the fundamental data structure in phrase-based models, and the training pipeline of most statistical machine translation (SMT) systems uses a word alignment for limiting the set of the suitable phrases in phrase extraction. Therefore, the accuracy of the phrase models are highly correlated with the word alignments quality, which are used to learn an accordance between the source and target words in parallel sentences. However, there is no theoretical support from the view of providing a formulation to describe the relationship between word alignments and machine translation performance.

We examine the Kazakh language, which is the majority language in the Republic of Kazakhstan. Kazakh is part of the Kipchak branch of the Turkic language family and part of the majority Ural-Altay family, in comparison with languages like English, is very rich in morphology.

The Kazakh language which words are generated by adding affixes to the root form is called an agglutinative language. We can derive a new word by adding an

V.-N. Huynh et al. (Eds.): IUKM 2015, LNAI 9376, pp. 326–335, 2015.
DOI: 10.1007/978-3-319-25135-6_31