Florian Daniel
Oscar Diaz (Eds.)

LNCS 9396

# Current Trends in Web Engineering

15th International Conference, ICWE 2015 Workshops
NLPIT, PEWET, SoWEMine
Rotterdam, The Netherlands, June 23–26, 2015
Revised Selected Papers

Springer

*Editors*
Florian Daniel
Università di Trento
Povo, Trento
Italy

Oscar Diaz
Universidad del Pais Vasco
San Sebastian
Spain

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-24799-1 ISBN 978-3-319-24800-4 (eBook)
DOI 10.1007/978-3-319-24800-4

Library of Congress Control Number: 2015950045

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

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)

# SMT: A Case Study of Kazakh-English Word Alignment

Amandyk Kartbayev(✉)

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

**Abstract.** In this paper, we present results from a set of experiments to determine the effect on translation quality, it depends on the particular kind of morphological preprocessing that can be represented by finite-state transducers. A high agglutinative nature of the Kazakh language under the condition of poor language resources makes an issue in the processing of derivational morphology. Our methods are focused on useful phrase pairs in word alignment and provide a most language independent approach, which may improve a translation into other morphological complex languages. We processed our algorithms over the Kazakh Wikipedia base of about 1.5 million unique lexeme and 230 million words overall. Our best translation system increases 3 BLEU points over the Kazakh-English baseline on a blind test set.

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

## 1 Introduction

In this work we focus on the word alignment process, which is the most important part of information recovery from a source with a lot of inflection. Particularly, we are interested in the sources where the given sentence pairs contain more new words with a less prior information about their nature. This is a challenging problem in machine translation and it is a hard to learn from the lexicon and usually repeats the similar errors again and again.

Morphological segmentation process intended to break words into morphemes, which are the basic semantic units and a key component for natural language processing systems. This is our current subtask in the machine translation project and we also desired to show that a simple segmentation scheme can perform pretty well as the most sophisticated one.

Most papers in statistical machine translation (SMT) oriented morphology analysis presents experiments that they consist of numerous experimentation to choose the best among a set of segmentation schemes. These morphological preprocessing schemes focused on various level of decomposition and compare the resulting translation performances, but usually use a subset of morphology and apply only a few simple rules in a segmentation process.


F. Daniel and O. Diaz (Eds.): ICWE 2015 Workshops, LNCS 9396, pp. 40–49, 2015.
DOI: 10.1007/978-3-319-24800-4_4

Florian Daniel · Oscar Diaz (Eds.)

# Current Trends in Web Engineering

15th International Conference, ICWE 2015 Workshops
NLPIT, PEWET, SoWEMine
Rotterdam, The Netherlands, June 23–26, 2015
Revised Selected Papers

Springer