SANDIA REPORT
SAND2012-10727
Unlimited Release
Printed December 2012

Discriminative Feature-Rich Models for Syntax-Based Machine Translation

Kevin R Dixon

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation,
a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's
National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online order:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2012-10727
Unlimited Release
Printed December 2012

Discriminative Feature-Rich Models for Syntax-Based Machine Translation

Kevin R. Dixon
Analytics & Cryptography
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185

Abstract

This report describes the campus executive LDRD "Discriminative Feature-Rich Models for Syntax-Based Machine Translation," which was an effort to foster a better relationship between Sandia and Carnegie Mellon University (CMU). The primary purpose of the LDRD was to fund the research of a promising graduate student at CMU; in this case, Kevin Gimpel was selected from the pool of candidates. This report gives a brief overview of Kevin Gimpel's research.

CONTENTS

1. INTRODUCTION

This report describes the progress of the campus executive LDRD “Discriminative Feature-Rich Models for Syntax-Based Machine Translation” which was an effort to foster a better relationship between Sandia and Carnegie Mellon University (CMU). The primary purpose of the LDRD was to fund the research of a promising graduate student at CMU; in this case, Kevin Gimpel was selected from the pool of candidates, and the funding ran in FY11 and FY12. Kevin Gimpel received his PhD from the Language Technologies Institute at CMU after successfully defending his thesis in summer 2012 and is now a research assistant professor at the Toyota Technological Institute at Chicago. This report will borrow from Kevin Gimpel’s PhD thesis [1] in describing his work

2. SPONSORED RESEARCH DESCRIPTION[1]

Fully-automated, high-quality machine translation promises to revolutionize human communication. But as anyone who has used a machine translation system knows, we are not there yet. In this thesis, we address four areas in which we believe translation quality can be improved across a large number of language pairs.

The first relates to flexible tree-to-tree translation modeling. Building translation systems for many language pairs requires addressing a wide range of translation divergence phenomena (Dorr, 1994). Recent research has shown clear improvement in translation quality by exploiting linguistic syntax for either the source or target language (Yamada and Knight, 2001; Galley et al., 2006; Zollmann and Venugopal, 2006; Liu et al., 2006). However, when using syntax for both languages ("tree-to-tree" translation), syntactic divergence hampers the extraction of useful rules (Ding and Palmer, 2005; Cowan et al., 2006; Ambati and Lavie, 2008; Liu et al., 2009a). Recent research shows that using soft constraints can substantially improve performance (Liu et al., 2009a; Chiang, 2010; Zhang et al., 2011; Hanneman and Lavie, 2011). Recently, Smith and Eisner (2006a) developed a flexible family of formalisms that they called quasi-synchronous grammar (QG). QG treats non-isomorphic structure softly using features rather than hard constraints. While a natural fit for syntactic translation modeling, the increased flexibility of the formalism has proved challenging for building real-world systems. In this thesis, we present the first machine translation system based on quasi-synchronous grammar.

Relatedly, we seek to unify disparate translation models. In designing a statistical model for translation, a researcher seeks to capture intuitions about how humans translate. This is typically done by specifying the form of translation rules and learning them automatically from large corpora.
The current trend is toward larger and increasingly-intricate rules. Some systems use rules with flat phrase mappings (Koehn et al., 2003), while others use rules inspired by linguistic syntax (Yamada and Knight, 2001). Neither is always better than the other (DeNeefe et al., 2007; Birch et al., 2009; Galley and Manning, 2010). In this thesis, we build a system that unifies rules from these two categories in a single model. Specifically, we use rules that combine phrases and dependency syntax by developing a new formalism called quasi-synchronous phrase dependency grammar.

In order to build these models, we need learning algorithms that can support feature-rich translation modeling. Due to characteristics of the translation problem, machine learning algorithms change when adapted to machine translation (Och and Ney, 2002; Liang et al., 2006a; Arun and Koehn, 2007; Watanabe et al., 2007; Chiang et al., 2008b), producing a breed of complex learning procedures that, though effective, are not well-understood or easily replicated. In this thesis, we contribute a new family of learning algorithms based on minimizing the structured ramp loss (Do et al., 2008). We develop novel variations on this loss, draw connections to several popular learning methods for machine translation, and develop algorithms for optimization. Our algorithms are effective in practice while remaining conceptually straightforward and easy to implement.

[1] Taken directly from Kevin Gimpel's PhD dissertation abstract.

Our final focus area is the use of syntactic structure for translation when linguistic annotations are not available. Syntax-based models typically use automatic parsers, which are built using corpora of manually-annotated parse trees. Such corpora are available for perhaps twenty languages (Marcus et al., 1993; Buchholz and Marsi, 2006; Petrov et al., 2012). In order to apply our models to the thousands of language pairs for which we do not have annotations, we turn to unsupervised parsers. These induce syntactic structures from raw text. The statistical NLP community has been doing unsupervised syntactic analysis for years (Magerman and Marcus, 1990; Brill and Marcus, 1992; Yuret, 1998; Paskin, 2002; Klein and Manning, 2002, 2004), but these systems have not yet found a foothold in translation research. In this thesis, we take the first steps in using unsupervised parsing for machine translation.

3. CONCLUSIONS

This report briefly describes the CMU campus executive LDRD "Discriminative Feature-Rich Models for Syntax-Based Machine Translation," which supported Kevin Gimpel's PhD research while he was at the Language Technology Institute at CMU.

4. REFERENCES

1. Kevin Gimpel, *Discriminative Feature-Rich Modeling for Syntax-Based Machine Translation*, PhD Thesis, Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, 2012. Available from http://ttic.uchicago.edu/~kgimpel/papers/gimpel_thesis.pdf

DISTRIBUTION

1	MS0899	Technical Library	9536 (electronic copy)
1	MS0359	D. Chavez, LDRD Office	1911

Sandia National Laboratories