SPRINGER BRIEFS IN STATISTICS

Florentina T. Hristea

The Naïve Bayes Model for Unsupervised Word Sense Disambiguation

Aspects Concerning Feature Selection

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Florentina T. Hristea

The Naïve Bayes Model for Unsupervised Word Sense Disambiguation

Aspects Concerning Feature Selection

Springer

Florentina T. Hristea
Faculty of Mathematics and Computer Science
Department of Computer Science
University of Bucharest
Bucharest
Romania

ISSN 2191-544X ISSN 2191-5458 (electronic)
ISBN 978-3-642-33692-8 ISBN 978-3-642-33693-5 (eBook)
DOI 10.1007/978-3-642-33693-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012949374

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To the memory of my father,
Prof. Dr. Theodor Hristea,
who has passed on to me
his love for words

Preface

The present work concentrates on the issue of feature selection for the Naïve Bayes model with application in unsupervised word sense disambiguation (WSD). It examines the process of feature selection while referring to an unsupervised corpus-based method for automatic WSD that relies on this specific statistical model. It concentrates on a distributional approach to unsupervised WSD based on monolingual corpora, with focus on the usage of the Naïve Bayes model as clustering technique.

While the Naïve Bayes model has been widely and successfully used in supervised WSD, its usage in unsupervised WSD has led to more modest disambiguation results and is less frequent. One could, in fact, say that it has been entirely dropped. The latest and most comprehensive survey[1] on WSD refers to the Naïve Bayes model strictly in conjunction with supervised WSD noting that "in spite of the independence assumption, the method compares well with other supervised methods" (Navigli 2009). It seems that the potential of this statistical model in unsupervised WSD continues to remain insufficiently explored. We feel that unsupervised WSD has not yet made full use of the Naïve Bayes model.

It is equally our belief that the Naïve Bayes model needs to be fed knowledge in order to perform well as clustering technique for unsupervised WSD. This knowledge can be fed in various ways and can be of various natures. The present work studies such knowledge of completely different types and hopes to initiate an open discussion concerning the nature of the knowledge that is best suited for the Naïve Bayes model when acting as clustering technique. Three different sources of such knowledge, which have been used only very recently in the literature (relatively to this specific clustering technique) are being examined and compared: WordNet, dependency relations, and web N-grams. This study ultimately concentrates not on WSD (which is regarded as an application) but on the issue of feeding knowledge to the Naïve Bayes model for feature selection.

[1] Navigli, R.: Word Sense Disambiguation: A Survey. ACM Comput. Surv. **41**(2), 1–69 (2009).

The present work represents a synthesis of 5 journal papers that have been authored or coauthored by us during the time interval 2008–2012, when our scientific interest was fully captured by the issue of feature selection for the Naïve Bayes model. This research is hereby extended, with two important additional conclusions being drawn in Chaps. 4 and 5. Each chapter will introduce knowledge of a different type, that is to be fed to the Naïve Bayes model, indicating those words (features) that should be part of the so-called "disambiguation vocabulary" when trying to decrease the number of parameters for unsupervised WSD based on this statistical model.

This work therefore places WSD with an underlying Naïve Bayes model at the border between unsupervised and knowledge-based techniques. It highlights the benefits of feeding knowledge (of various natures) to a knowledge-lean algorithm for unsupervised WSD that uses the Naïve Bayes model as clustering technique.

Our study will show that a basic, simple knowledge-lean disambiguation algorithm, hereby represented by the Naïve Bayes model, can perform quite well when provided knowledge in an appropriate way. It will equally justify our belief that the Naïve Bayes model still holds a promise for the open problem of unsupervised WSD.

Toulouse, France, November 2011 Florentina T. Hristea

Acknowledgments

The author expresses her deepest gratitude to Professor Ted Pedersen for having provided the dataset necessary for performing the presented tests and comparisons with respect to adjectives and verbs. We are equally indebted to two anonymous referees for their valuable comments and suggestions. This research was supported by the National University Research Council of Romania (the "Ideas" research program, PN II—IDEI), Contract No. 659/2009.