

Eraldo Luís Rezende Fernandes

**Entropy Guided Feature Generation for
Structure Learning**

Tese de Doutorado

Thesis presented to the Programa de Pós-Graduação em
Informática, of the Departamento de Informática do Centro
Técnico Científico da PUC-Rio, as partial fulfillment of the
requirements for the degree of Doutor.

Advisor: Prof. Ruy Luiz Milidiú

Rio de Janeiro
September 2012

Eraldo Luís Rezende Fernandes

Entropy Guided Feature Generation for Structure Learning

Thesis presented to the Programa de Pós-Graduação em
Informática, of the Departamento de Informática do Centro
Técnico Científico da PUC-Rio, as partial fulfillment of the
requirements for the degree of Doutor.

Prof. Ruy Luiz Milidiú

Advisor

Departamento de Informática — PUC-Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão

Departamento de Informática — PUC-Rio

Prof. Valmir Carneiro Barbosa

UFRJ

Prof. Cícero Nogueira dos Santos

IBM Research

Prof. Daniel Schwabe

Departamento de Informática — PUC-Rio

Prof. José Eugenio Leal

Coordinator of the Centro Técnico Científico — PUC-Rio

Rio de Janeiro — September 6, 2012

All rights reserved.

Eraldo Luís Rezende Fernandes

Graduated from the Universidade Federal de Mato Grosso do Sul in Ciência da Computação and obtained a degree of Mestre em Informática at PUC-Rio. He is a lecturer and a researcher at the Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

Bibliographic data

Fernandes, Eraldo Luís Rezende

Entropy Guided Feature Generation for Structure Learning
/ Eraldo Luís Rezende Fernandes ; advisor: Ruy Luiz Milidiú.
— 2012.

93 f. : il. ; 30 cm

Tese (Doutorado em Informática)-Pontifícia Universidade
Católica do Rio de Janeiro, Rio de Janeiro, 2012.

Inclui bibliografia

1. Informática – Teses. 2. Aprendizado de Estruturas.
3. Geração de Atributos. 4. Entropia. 5. Processamento
de Linguagem Natural. I. Milidiú, Ruy Luiz. II. Pontifícia
Universidade Católica do Rio de Janeiro. Departamento de
Informática. III. Título.

CDD: 004

Acknowledgments

First of all, I am thankful for all my family which has always supported me. We have had amazing moments during so many meetings, and it is always my pleasure to be with them. Specially, I would like to thank my parents, Eraldo and Sônia, my sister, Luciana, and my little brother, Lúciu. They constitute my ultimate foundation for all aspects of my life.

I am unable to express in words all my love and gratitude for my dear wife, Valéria. She stood by me during the most difficult and sad moments of my PhD course. And, more importantly, we are always enjoying each other and having fun with the simplest things.

I also want to thank my sisters-in-law, Vanessa and Patrícia, and my brother-in-law, Márcio.

I will never forget the great moments I have spent in the lab. Thank you very much, all *Learnlings*. Specially, I am glad for having closely worked with Carlos, Cícero, Eduardo, Leandro and William.

I am thankful to my advisor, Ruy, who have taught me so many things regarding different aspects of my life.

I thank all my friends from Barcelona, where I spent one great year at Yahoo! Research Lab. Specially, I need to thank my supervisor and friend, Ulf, and all the *Berlineros*.

Finally, I need to thank my colleagues from IFG for all the support. Specially, José Antônio, his family and Sérgio Henrique.

Abstract

Fernandes, Eraldo Luís Rezende; Milidiú, Ruy Luiz. **Entropy Guided Feature Generation for Structure Learning**. Rio de Janeiro, 2012. 93p. DSc Thesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Structure learning consists in learning a mapping from inputs to structured outputs by means of a sample of correct input-output pairs. Many important problems fit into this setting. Natural language processing provides several tasks that can be formulated and solved as structure learning problems. Dependency parsing, for instance, involves the prediction of a tree underlying a sentence. Feature generation is an important subtask of structure learning which, usually, is partially solved by a domain expert that builds complex discriminative feature templates by conjoining the available basic features. This is a limited and expensive way to generate features and is recognized as a modeling bottleneck.

In this work, we propose an automatic feature generation method for structure learning problems. This method is entropy guided since it generates complex features based on the conditional entropy of local output variables given the available input features. We experimentally compare the proposed method with two important alternative feature generation methods, namely manual template generation and polynomial kernel methods. Our experimental findings indicate that the proposed method is more attractive than both alternatives. It is much cheaper than manual templates and computationally faster than kernel methods. Additionally, it is simpler to control its generalization performance than with kernel methods.

We evaluate our method on nine datasets involving five natural language processing tasks and four languages. The resulting systems present state-of-the-art comparable performances and, particularly on part-of-speech tagging, text chunking, quotation extraction and coreference resolution, remarkably achieve the best known performances on different languages like Arabic, Chinese, English, and Portuguese. Furthermore, our coreference resolution systems achieve the very first place on the Conference on Computational Natural Language Learning 2012 Shared Task. The competing systems were ranked by the mean score over three languages: Arabic, Chinese and English. Our approach obtained the best performances among all competitors for all the three languages.

Our feature generation method naturally extends the general structure learning framework and is not restricted to natural language processing tasks.

Keywords

Structure Learning. Feature Generation. Entropy. Natural
Language Processing.

Resumo

Fernandes, Eraldo Luís Rezende; Milidiú, Ruy Luiz. **Geração de Atributos Guiada por Entropia para Aprendizado de Estruturas**. Rio de Janeiro, 2012. 93p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Aprendizado de estruturas consiste em aprender um mapeamento de variáveis de entrada para saídas estruturadas a partir de exemplos de pares entrada-saída. Vários problemas importantes podem ser modelados desta maneira. O processamento de linguagem natural provê diversas tarefas que podem ser formuladas e solucionadas através do aprendizado de estruturas. Por exemplo, parsing de dependência envolve o reconhecimento de uma *árvore* implícita em uma frase. Geração de atributos é uma sub-tarefa importante do aprendizado de estruturas. Geralmente, esta sub-tarefa é realizada por um especialista que constrói gabaritos de atributos complexos e discriminativos através da combinação dos atributos básicos disponíveis na entrada. Esta é uma forma limitada e cara para geração de atributos e é reconhecida como um gargalo de modelagem.

Neste trabalho, propomos um método automático para geração de atributos para problemas de aprendizado de estruturas. Este método é *guiado por entropia* já que é baseado na entropia condicional de variáveis locais de saída dados os atributos básicos. Comparamos experimentalmente o método proposto com dois métodos alternativos para geração de atributos: geração manual e métodos de kernel polinomial. Nossos resultados mostram que o método de geração de atributos guiado por entropia é superior aos dois métodos alternativos em diferentes aspectos. Nosso método é muito mais barato do que o método manual e computacionalmente mais rápido que o método baseado em kernel. Adicionalmente, ele permite o controle do seu poder de generalização mais facilmente do que métodos de kernel.

Nós avaliamos nosso método em nove datasets envolvendo cinco tarefas de linguística computacional e quatro idiomas. Os sistemas desenvolvidos apresentam resultados comparáveis aos melhores sistemas atualmente e, particularmente para etiquetagem morfossintática, identificação de sintagmas, extração de citações e resolução de coreferência, obtêm os melhores resultados conhecidos para diferentes idiomas como Árabe, Chinês, Inglês e Português. Adicionalmente, nosso sistema de resolução de coreferência obteve o primeiro lugar na competição *Conference on Computational Natural Language Learning 2012 Shared Task*. O sistema vencedor foi determinado pela média de desempenho em três idiomas: Árabe, Chinês e Inglês. Nosso sistema obteve o melhor desempenho nos três idiomas avaliados.

Nosso método de geração de atributos estende naturalmente o framework de aprendizado de estruturas e não está restrito a tarefas de processamento de linguagem natural.

Palavras-chave

Aprendizado de Estruturas. Geração de Atributos. Entropia.
Processamento de Linguagem Natural.

Contents

1	Introduction	13
1.1	Ad Hoc Approaches	13
1.2	Linear Discriminative Models	15
1.3	Nonlinearity	20
1.4	Entropy-Guided Structure Learning	21
1.5	Contributions	22
1.6	Dissertation Organization	24
2	Structure Learning Framework	26
2.1	Dependency Parsing	26
2.2	Large Margin Training	28
2.3	Latent Structure Training	29
2.4	Empirical Results	31
3	Entropy-Guided Feature Generation	33
3.1	Basic Dataset	34
3.2	Conditional Entropy and Information Gain	34
3.3	Decision Tree Learning	36
3.4	Feature Templates	37
3.5	Generated Features	38
3.6	Empirical Results	38
4	Entropy-Guided Structure Learning Framework	40
4.1	Feature Factorization	40
4.2	Entropy-Guided Feature Generation	41
4.3	Training Algorithm	42
4.4	Kernelization	44
4.5	Empirical results	46
5	Prediction Problems	47
5.1	Rooted Tree	48
5.2	Sequence Labeling	49
5.3	Sequence Segmentation	49
5.4	Clustering	50
6	Dependency Parsing	51
6.1	Task Formalization	51
6.2	Feature Factorization	51
6.3	Prediction Problem	51
6.4	Basic Features	52
6.5	Empirical Results	53
7	Part-of-Speech Tagging	55
7.1	Task Formalization	55
7.2	Feature Factorization	56

7.3	Prediction Problem	57
7.4	Basic Features	58
7.5	Empirical Results	59
8	Text Chunking	61
8.1	Task Formalization	61
8.2	Feature Factorization	62
8.3	Prediction Problem	62
8.4	Basic Features	62
8.5	Empirical Results	62
9	Quotation Extraction	65
9.1	Task Formalization	65
9.2	Feature Factorization	66
9.3	Prediction Problem	66
9.4	Basic Features	67
9.5	Empirical Results	67
10	Coreference Resolution	69
10.1	Task Formalization	70
10.2	Feature Factorization	70
10.3	Prediction Problem	73
10.4	Basic Features	74
10.5	Data Preparation	74
10.6	Empirical Results	78
11	Conclusions	84

List of Figures

1.1	Dependency tree example.	14
1.2	Binary perceptron algorithm.	16
1.3	Multiclass perceptron algorithm.	17
1.4	Generalized perceptron algorithm for dependency parsing.	18
1.5	Structure perceptron algorithm.	19
1.6	Averaged structure perceptron algorithm.	19
1.7	Entropy-Guided Structure Learning framework.	22
2.1	Dependency tree represented as arcs (\mathbf{y}) and as head vector.	26
2.2	Large margin structure perceptron algorithm for dependency parsing.	29
2.3	Latent structure perceptron algorithm.	31
3.1	Entropy $H(y)$ of a random binary variable y versus $Pr[y = 1]$ that is denoted by p .	35
3.2	A decision tree.	36
3.3	Feature template induction from a decision tree.	37
4.1	ESL training algorithm – the entropy-guided large-margin structure perceptron.	43
7.1	Part-of-speech tagging example.	55
7.2	Illustrative directed acyclic graph for a sentence $\mathbf{x} = (x_1, x_2, x_3)$ and a tagset $S = \{a, b\}$. The continuous path $(y_{1,b}, y_{2,a}, y_{3,a})$ corresponds to the labeling $\mathbf{y} = (b, a, a)$.	57
8.1	Text chunking example.	61
9.1	Quotation extraction example.	65
10.1	Document with nine highlighted mentions that refer to three different entities: North Korea is referenced by mentions $\{a_1, a_2, a_3, a_4\}$; the U.S. is referenced by $\{b_1, b_2\}$; and Madeleine Albright by $\{c_1, c_2, c_3\}$. The letter in the mention subscript identifies its entity cluster and the number uniquely identifies the mention within its cluster.	69
10.2	Coreference tree for the cluster a in Figure 10.1.	71
10.3	Document tree with three coreference trees that corresponds to the text in Figure 10.1. Dashed lines indicate artificial arcs.	72
10.4	Latent structure perceptron algorithm.	73

List of Tables

1.1	Comparison of EFG with other feature generation methods.	21
1.2	Comparison of ESL with state-of-the-art systems.	22
2.1	Performances of dependency parsers using manual templates on the Portuguese CoNLL-2006 dataset. These systems use different learning algorithms and also different basic features.	32
3.1	Basic dataset for the sentence in Figure 2.1.	34
3.2	Performances of EFG and manual templates on the Portuguese CoNLL-2006 dependency parsing dataset.	39
4.1	Comparison of ESL to second-degree polynomial kernel.	46
5.1	List of tasks and the corresponding output structures and prediction problems.	47
6.1	Bosque dependency parsing dataset statistics.	53
6.2	Performances of ESL and state-of-the-art systems on the Portuguese CoNLL-2006 dependency parsing dataset.	53
7.1	Basic statistics of the part-of-speech tagging datasets.	59
7.2	Performances on the Mac-Morpho dataset.	59
7.3	Performances on the Brown dataset.	60
8.1	Basic statistics of the text chunking datasets.	62
8.2	Performances on the Bosque dataset.	63
8.3	Performances on the CoNLL-2000 dataset.	63
9.1	GloboNotes dataset statistics.	67
9.2	Performances on the GloboQuotes dataset.	68
10.1	Description of all 70 basic features.	75
10.2	State-of-the-art systems for multilingual unrestricted coreference resolution in OntoNotes. Performances on the CoNLL-2012 Shared Task test sets.	79
10.3	Detailed performance of our system on the CoNLL-2012 Shared Task test sets.	79
10.4	EFG effect on system performance for the English development set.	80
10.5	Root loss value effect on development set performances.	81
10.6	Effect whether nested coreferring mentions are considered or not for the Chinese language.	82
10.7	Supplementary results on the test sets with different configurations (Config) for parse quality and mention candidates (parse/mentions). Parse quality can be automatic (A) or golden (G); and mention candidates can be automatically identified (A), golden mention boundaries (GB) or golden mentions (GM).	82