

Markus Stengel

Strings of Natural Languages

*Unsupervised Analysis and Segmentation on the
Expression Level*

Bibliographic information published by the German National Library:

The German National Library lists this publication in the National Bibliography; detailed bibliographic data are available on the Internet at <http://dnb.dnb.de> .

This book is copyright material and must not be copied, reproduced, transferred, distributed, leased, licensed or publicly performed or used in any way except as specifically permitted in writing by the publishers, as allowed under the terms and conditions under which it was purchased or as strictly permitted by applicable copyright law. Any unauthorized distribution or use of this text may be a direct infringement of the author s and publisher s rights and those responsible may be liable in law accordingly.

Copyright © 2006 Diplomica Verlag GmbH
ISBN: 9783836606271

Markus Stengel

Strings of Natural Languages

Unsupervised Analysis and Segmentation on the Expression Level

Markus Stengel

Strings of Natural Languages

*Unsupervised Analysis and Segmentation on the
Expression Level*

Markus Stengel

Strings of Natural Languages

Unsupervised Analysis and Segmentation on the Expression Level

ISBN: 978-3-8366-0627-1

Druck Diplomica® Verlag GmbH, Hamburg, 2008

Zugl. Eberhard-Karls-Universität Tübingen, Tübingen, Deutschland, Diplomarbeit, 2006

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtes.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden, und die Diplomarbeiten Agentur, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

© Diplomica Verlag GmbH

<http://www.diplomica.de>, Hamburg 2008

Printed in Germany

CONTENTS

Table of Contents	iii
List of Figures	vi
List of Tables	viii
List of Algorithms	ix
List of Abbreviations	xi
Introduction	1
1 Language	3
1.1 Definitions	3
1.2 Languages	5
1.2.1 English	6
1.2.2 German	10
1.2.3 Hebrew	13
1.2.4 Japanese	14
2 Categorization	23
2.1 Definitions	23
2.2 Sample Application	24

2.3	Conclusion	26
3	Analysis Methods and Techniques	27
3.1	Level of Abstraction	27
3.2	Data Compression	27
3.2.1	Overview	27
3.2.2	Information content and its quantification	28
3.2.3	Kinds of data compression	30
3.2.4	Run length encoding	31
3.2.5	Dictionary-based data compression: LZ78, LZW, LZMW	33
3.2.6	LZMW78	37
3.2.7	Sample application	39
3.3	Longest Common Subsequence	40
3.3.1	Overview	40
3.3.2	Application	41
3.4	Statistics: N-Gram and Term Frequency	43
3.4.1	Definitions	43
3.4.2	Limited applicability of published statistics	44
3.4.3	The challenges of collecting statistics	45
3.4.4	Fixed term size	46
3.4.5	Variable term size	48
3.4.6	Suffix tree	48
3.4.7	Suffix array	52
3.5	Cryptology	56
3.5.1	Motivation	56
3.5.2	Character frequency	57
3.5.3	Index of coincidence	58
3.5.4	Patterns	59
4	Tasks and Results	61
4.1	Experimental Setup	61
4.1.1	Corpora	61
4.1.2	Preprocessing	63

LIST OF FIGURES

3.1	Types of data compression	31
3.2	Suffix tree examples	50
3.3	Suffix array structures after sorting	54
4.1	Entropy images for the corpora G1, H1 and J3	68
4.2	Index of coincidence difference plots for the corpora E1, J4, G4 and G5	70
5.1	The chain of tools developed in this work	98
A.1	Language tree	107
B.1	Index of coincidence plot: G1	109
B.2	Index of coincidence plot: G1	109
B.3	Index of coincidence plot: G1	110
B.4	Index of coincidence plot: G1	110
B.5	Index of coincidence plot: G2	111
B.6	Index of coincidence plot: G3	111
B.7	Index of coincidence plot: G4	112
B.8	Index of coincidence plot: G5	112
B.9	Index of coincidence plot: H1	113
B.10	Index of coincidence plot: H2	113
B.11	Index of coincidence plot: J1	114
B.12	Index of coincidence plot: J2	114

B.13 Index of coincidence plot: J3	115
B.14 Index of coincidence plot: J4	115
B.15 Index of coincidence plot: J5	116
B.16 Suffixes, prefixes and reduced SUs for E3	119
B.17 Suffixes, prefixes and reduced SUs for G3	120
B.18 Suffixes, prefixes and reduced SUs for H1	121
B.19 Suffixes, prefixes and reduced SUs for H2	122
B.20 Suffixes, prefixes and reduced SUs for J3	123
B.21 Segmentation results for H1 and H2	124
B.22 Segmentation results for G3	124

LIST OF TABLES

1.1	Language families and languages	5
1.2	Frequencies of constituent orders	6
2.1	Sample tokenization categories	24
2.2	Sample text tokenization	25
2.3	Results of sample text categorization	25
2.4	Dictionary built from sample text categorization	25
2.5	Sample text reencoded to the dictionary built from categorization	26
3.1	Sample encoding with LZ78	34
3.2	Dictionary of an LZW sample compression	35
3.3	Encoding steps of an LZW sample compression	36
3.4	Dictionary of an LZMW sample compression	36
3.5	Sample encoding with LZMW78	38
3.6	LZMW78 dictionary after encoding the vector (3,1,4,1,0,1,5,2)	39
3.7	LCS examples	41
3.8	Various categorizations exploitable by LCS	42
3.9	Frequencies of English and German letters	45
3.10	Possible terms dependent on alphabet size	46
3.11	Prefixes and suffixes of the string ‘BANANA\$’	49
3.12	Suffix array illustration	53
3.13	Suffix array and prefix tables	54

3.14	Order of letter frequency	57
3.15	Index of coincidence for various languages	58
4.1	Corpora used in this work	62
4.2	Experimental Setup: system specifications and implementation	64
4.3	Sample rating results and rankings	65
4.4	Sample meta-rating results	66
4.5	Maximum ratio of IC differences to IC for the individual corpora	71
4.6	Character order of the corpora	72
4.7	Pangram-ending character order	74
4.8	LCS and compression results: syntactic separators	78
4.9	Sample problem dictionary for prefix tables	80
4.10	Aligner results for biblical corpora	81
4.11	Suffixes, prefixes and reduced SUs for biblical corpora	85
4.12	Suffixes, prefixes and reduced SUs for J3	87
4.13	compound detection results for biblical corpora	90
4.14	Most frequent meta-rated terms	94
A.1	hiragana	104
A.2	katakana	105
A.3	Hebrew alphabet	106
B.1	Aligner results for various corpora	117
B.2	Aligner results for various corpora	118
B.3	Detected compound samples for E3	125
B.4	Detected compound samples for G3	126
B.5	Detected compound samples for H1	127
B.6	Detected compound samples for H2	128
B.7	Detected compound samples for J3	129

LIST OF ALGORITHMS

1	Entropy image creation	67
2	Index of coincidence computation	69
3	Detection of syntactic separators with LCS and LZMW78	76
4	Detect syntactic separators by aligning at selected strings	80
5	Detection of prefixes and suffixes	83
6	Detect and split compounds	89

List of Abbreviations

AIC	algorithmic information content
IC	index of coincidence
ID	identification (number)
KCC	Kolmogorov-Chaitin complexity
MR	meta-rating
PSR	prefix, suffix, and/or reduced segmentation unit
RLE	run-length encoding
SIC	Shannon information content
SOV	subject-object-verb (sentence structure)
SU	segmentation unit
SVO	subject-verb-object (sentence structure)
TF	term frequency
LCS	longest common subsequence
LZW	Lempel-Ziv-Welch (compression)
LZMW	Lempel-Ziv-Miller-Wegman (compression)
LZMW78	my modification of LZMW (compression)

Introduction

The limits of my language are the limits of my mind. All I know is what I have words for. – Ludwig Wittgenstein

Everyone who has ever learned a second language knows how hard it is. There are always differences: Some are glaringly obvious, and others are so subtle that even their concepts are difficult to understand. One major reason for this is the way we learn: We try to translate the words and concepts of the other language into those of our own language which we are comfortable with.

As long as the languages are fairly similar, this works quite well. However, when the languages differ to a great degree, problems are bound to appear. For example, to someone whose first language is French, English is not difficult to learn. In fact, he can pick up any English book and at the very least recognize words and sentences. But if he is tasked with reading a Japanese text, he will be completely lost: No familiar letters, no whitespace, and only occasionally a glyph that looks similar to a punctuation mark appears.

Nevertheless, anyone can learn any language. Correct pronunciation and understanding alien utterances may be hard for the individual, but as soon as the words are transcribed to some kind of script, they can be studied and - given some time - understood. The script thus offers itself as a reliable medium of communication.

Sometimes the script can be very complex, though. For instance, the Japanese language is not much more difficult than German - but the Japanese script is. If someone untrained in the language is given a Japanese book and told to create a list of its vocabulary, he will most likely have to succumb to the task.

Or does he not? Are there maybe ways to analyze the text, regardless of his unfamiliarity with this type of script and language? Should there not be characteristics shared by all languages which can be exploited?