

การแก้คำผิดแบบไม่ตั้งใจโดยอัตโนมัติในภาษาไทย เพื่อการสื่อสารกับหุ่นยนต์สนทนา



นางสาววนิดา เกษรสุวรรณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-53-2846-4

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

T2247786X

AUTOMATIC CORRECTION FOR UNINTENTIONAL TYPING ERRORS
IN THAI LANGUAGE FOR COMMUNICATION WITH CHAT ROBOT

Ms. Wanida Kessuwon



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2005

ISBN 974-53-2846-4

481859

นางสาววนิดา เกษรสุวรรณ : การแก้คำผิดแบบไม่ตั้งใจโดยอัตโนมัติในภาษาไทย เพื่อการสื่อสารกับหุ่นยนต์สนทนา. (AUTOMATIC CORRECTION FOR UNINTENTIONAL TYPING ERRORS IN THAI LANGUAGE FOR COMMUNICATION WITH CHAT ROBOT) อ. ที่ปรึกษา : อาจารย์ ดร.วิษณุ โคตรจรัส, 82 หน้า. ISBN 974-53-2846-4.

วิทยานิพนธ์นี้นำเสนอวิธีแก้ไขคำผิดแบบไม่ตั้งใจโดยอัตโนมัติในภาษาไทย เพื่อนำไปใช้กับหุ่นยนต์สนทนา ซึ่งใช้คุณสมบัติอักขระข้างเคียงบนแป้นพิมพ์ช่วยในการแก้ไขคำผิด โดยได้มีการทดสอบเพื่อสังเกตพฤติกรรมในการพิมพ์ พบรูปแบบการพิมพ์ผิดในภาษาไทยว่า 93.54 เปอร์เซ็นต์จากคำผิดที่พบทั้งหมด เกิดจากความผิดพลาดทั้ง 4 กรณีประกอบกัน คือ แทนที่ เกิน ตก สลับ (เรียงตามลำดับปริมาณที่พบจากมากไปน้อย) และตำแหน่งอักขระที่ผิดนั้นเฉลี่ยอยู่ตำแหน่งที่ 58.36 เปอร์เซ็นต์ของความยาวคำ

ในงานวิจัยนี้นำรูปแบบการพิมพ์ผิดที่ได้มาออกแบบอัลกอริทึมแก้ไขคำผิด แล้วทำการทดสอบประสิทธิภาพด้วยบทสนทนาที่มีคำผิดแบบไม่ตั้งใจทั้งหมดจำนวน 120 ประโยค (ยกเว้นความผิดพลาดที่มาจากการพิมพ์ตก) พบว่าหุ่นยนต์สามารถตอบได้คิดเป็น 95 เปอร์เซ็นต์ ซึ่งแสดงให้เห็นว่าหุ่นยนต์สามารถทำงานได้ประสิทธิภาพมากขึ้นหากใช้อัลกอริทึมนี้

ภาควิชา.... วิศวกรรมคอมพิวเตอร์.....ลายมือชื่อนิสิต.....
 สาขาวิชา....วิทยาศาสตร์คอมพิวเตอร์.....ลายมือชื่ออาจารย์ที่ปรึกษา.....
 ปีการศึกษา2548.....



46704653 : MAJOR Computer Science

KEY WORD: CHAT ROBOT / PATTERN MATCHING / AIML / AUTOMATIC CORRECTION / UNINTENTIONAL TYPING ERRORS / WORD SEGMENTATION

WANIDA KESSUWON : AUTOMATIC CORRECTION FOR UNINTENTIONAL TYPING ERRORS IN THAI LANGUAGE FOR COMMUNICATION WITH CHAT ROBOT. THESIS ADVISOR : VISHNU KOTRAJARAS, Ph.D., 82 pp. ISBN 974-53-2846-4.

This thesis proposes an algorithm, which is an automatic correction algorithm for unintentional typing errors in Thai language, for communicating with chat robot. It uses the characteristic of adjacent alphabets on keyboard for error correction. Our investigation found that 93.54 percent of all misspelled words in Thai language contains four kinds of typing errors : substitution, insertion, deletion and transposition (arranged from maximum to minimum frequency). Average error position in misspelled words is at 58.36 percent of the word length.

This thesis uses the discovered typing error patterns to design an error correction algorithm. The algorithm is tested on 120 sentences with unintentional typing errors (except error from deletion typing). The chat robot is able to identify 95 percent of the errors. It shows that the chat robot can work more effectively if it uses this algorithm.

Department.... Computer Engineering.... Student's signature.....
Field of study.... Computer Science..... Advisor's signature.....
Academic year ...2005.....

Wanida Kessuwon

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จได้ด้วยความกรุณาของอาจารย์ ดร.วิษณุ โคตรจรัส อาจารย์ที่ปรึกษา ซึ่งเป็นผู้ให้ความรู้ คำปรึกษา และข้อคิดเห็นอันเป็นประโยชน์ต่องานวิจัย รวมทั้งให้โอกาส ความช่วยเหลือ และกำลังใจที่ดีแก่ผู้วิจัยเสมอมา

ขอขอบคุณคณะกรรมการสอบวิทยานิพนธ์ ที่ได้กรุณาให้คำแนะนำในการแก้ไข วิทยานิพนธ์ให้มีคุณภาพยิ่งขึ้น

ขอขอบคุณอาจารย์ ดร.อรรถวิทย์ สุดแสง ที่กรุณาให้ใช้ห้องปฏิบัติการ ISL2 เป็นเวลา 5 เดือน และได้ให้ข้อคิดเห็น และคำแนะนำอันเป็นประโยชน์เกี่ยวกับการหาสถิติการพิมพ์ผิดใน ภาษาไทยของงานวิจัยนี้

ขอขอบคุณอาจารย์ ดร.โปรดปราน บุญยพุกกณะ และอาจารย์ ดร.อดิวงส์ สุชาโต ที่ให้ความกรุณาแก่ผู้วิจัยได้เป็นส่วนหนึ่งของห้องปฏิบัติการ SLS เป็นเวลา 9 เดือน

ขอขอบคุณอาจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์ ที่ให้ความช่วยเหลือในการประสานงานกับ เนคเทคเกี่ยวกับข้อมูลของโปรแกรมตัดคำภาษาไทย Swath และให้คำปรึกษาด้านต่างๆ เสมอมา

ขอขอบคุณอาจารย์ฐานิศา เกียรติบริวารมี ที่คอยให้คำปรึกษาในยามที่ข้าพเจ้าท้อแท้ และให้กำลังใจที่ดีเสมอ

ขอขอบคุณอาจารย์ เพื่อนๆ พี่ๆ และน้องๆ ในภาควิชาที่ช่วยทำแบบทดสอบในการเก็บ สถิติการพิมพ์ผิดในภาษาไทย และได้ให้ข้อความบทรสนทนาบางส่วนของตนเอง เพื่อให้ผู้วิจัย นำมาใช้เป็นตัวอย่างข้อมูลในการออกแบบและทดสอบงานวิจัย

ขอบคุณน้องกวิน อุชุกานนท์ชัย ที่ให้คำปรึกษาและความช่วยเหลือในการเขียนโปรแกรม ด้วยภาษาจาวา พร้อมทั้งให้โค้ดตัวอย่างการทำผลคูณคาร์ทีเซียน

ขอบคุณสมาชิกห้องปฏิบัติการชั้น 20 และห้องปฏิบัติการอื่นๆ ที่คอยช่วยเหลือ ให้ คำปรึกษา และให้กำลังใจในการทำวิจัยนี้

ท้ายที่สุด ผู้วิจัยใคร่ขอขอบพระคุณ คุณพ่อ คุณแม่ และญาติพี่น้องทุกคนในครอบครัว ที่ คอยดูแลห่วงใย เป็นกำลังใจ และให้การสนับสนุนในทุกๆ ด้านตลอดมา

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ	ช
สารบัญตาราง.....	ฅ
สารบัญภาพ.....	ญ
บทที่	
1 บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา.....	1
1.2 วัตถุประสงค์ของการวิจัย.....	2
1.3 ขอบเขตของการวิจัย	2
1.4 ขั้นตอนและวิธีดำเนินการวิจัย.....	3
1.5 ประโยชน์ที่คาดว่าจะได้รับการวิจัย	3
1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์.....	4
2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง.....	5
2.1 การตัดคำภาษาไทย	5
2.1.1 การตัดคำภาษาไทยโดยใช้คุณลักษณะ.....	5
2.1.2 โปรแกรมตัดคำภาษาไทยเวิร์ดคัท	6
2.2 การตรวจสอบและแก้ไขคำผิด	8
2.2.1 การตรวจสอบส่วนที่ไม่เป็นคำ.....	9
2.2.2 การแก้ไขคำผิดโดยไม่คำนึงถึงสภาพรอบข้าง.....	9
2.2.3 การแก้ไขคำผิดโดยขึ้นกับรูปประโยค.....	10
2.3 หุ่นยนต์สนทนา.....	10
2.4 เอไอเอ็มแอล.....	10
3 การทดลองหาสถิติการพิมพ์ผิดในภาษาไทย	19
3.1 การออกแบบการทดลองหาสถิติการพิมพ์ผิดในภาษาไทย.....	19
3.2 การวิเคราะห์ผลการทดลอง	21
3.2.1 การหาสัดส่วนความผิดพลาดในการพิมพ์.....	21
3.2.2 การหาตำแหน่งอักขระต้นเหตุของคำผิด	25

บทที่	หน้า
4 การออกแบบวิธีการแก้คำผิดอัตโนมัติสำหรับหุ่นยนต์สนทนา	30
4.1 วิธีการแก้คำผิด	30
4.1.1 การแทนที่	33
4.1.1.1 การหาส่วนที่ไม่เป็นคำ	33
4.1.1.2 การสร้างรายการใกล้เคียงคำผิด	33
4.1.2 การเกิน	35
4.1.3 การสลับ	36
4.2 ส่วนตัวแปลภาษาสำหรับเอกสารเอไอเอ็มแอล	36
4.3 การแก้คำผิดในกรณีที่มีคำผิดหลายทีในประโยค	38
4.4 การทดสอบประสิทธิภาพอัลกอริทึมแก้คำผิดกับงานอื่นในภาษาไทย	38
4.4.1 โปรแกรมไมโครซอฟท์เวิร์ดเอ็กซ์พี	38
4.4.2 โปรแกรมตรวจสอบตัวสะกดภาษาไทย	40
5 การทดสอบประสิทธิภาพการสื่อสารกับหุ่นยนต์สนทนา	42
5.1 การเลือกข้อความที่นำมาทดสอบ	42
5.2 การสร้างฐานความรู้ให้กับหุ่นยนต์สนทนา	43
5.3 การทดสอบสื่อสารกับหุ่นยนต์สนทนา	44
5.4 ผลการทดสอบประสิทธิภาพการสื่อสารกับหุ่นยนต์สนทนา	46
6 สรุปผลการวิจัยและข้อเสนอแนะ	48
6.1 สรุปผลการวิจัย	48
6.2 ปัญหาและข้อจำกัดที่พบจากการวิจัย	49
6.3 ข้อเสนอแนะ	50
รายการอ้างอิง	51
ภาคผนวก	53
ภาคผนวก ก แบบทดสอบที่ใช้ในการเก็บสถิติการพิมพ์ผิดภาษาไทย	54
ภาคผนวก ข ประโยคสนทนาที่มีคำผิดจำนวน 120 ประโยค	74
ภาคผนวก ค ประโยคสนทนาที่นำไปสร้างแพทเทิร์นจำนวน 120 ประโยค	78
ประวัติผู้เขียนวิทยานิพนธ์	82

สารบัญตาราง

ตารางที่	หน้า
3.1 สรุปจำนวนความผิดพลาดแบบต่างๆ ของการพิมพ์	22
3.2 สรุปสัดส่วนความผิดพลาดของคำผิด	24
3.3 สรุปตำแหน่งอักขระต้นเหตุของคำผิด	26
4.1 เปรียบเทียบผลการทดลองกับไมโครซอฟท์เวิร์ดเอ็กซ์พี	39
4.2 เปรียบเทียบผลการทดลองกับงานวิจัยของทิวา	40

สารบัญภาพ

ภาพที่	หน้า
2.1 กราฟแสดงการเพิ่มเส้นเชื่อมเพื่อแทนคำที่ไม่รู้จัก.....	7
2.2 กราฟแสดงการเพิ่มเส้นเชื่อมเพื่อแทนคำที่มีในพจนานุกรม.....	7
2.3 กฎในการจับกลุ่มอักขระ	7
2.4 กราฟแสดงการเพิ่มเส้นเชื่อมเพื่อแทนกลุ่มอักขระตามกฎการจับกลุ่มอักขระ	8
2.5 กราฟวิธีที่เป็นไปได้ในการตัดคำ	8
3.1 การแบ่งกลุ่มของคำที่พิมพ์ต่างไปจากต้นฉบับ	24
3.2 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบ่งตามความยาวคำ	27
3.3 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์แทนที่	28
3.4 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์เกิน.....	28
3.5 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์สลับ	28
3.6 กราฟแสดงความหนาแน่นของตำแหน่งอักขระต้นเหตุคำผิดแบบพิมพ์ตก	29
4.1 ตัวอย่างลำดับตำแหน่งในการแก้ไขคำผิด.....	31
4.2 อัลกอริทึมในการแก้ไขคำผิด.....	32
4.3 ผังอักขระที่อยู่ประชิดตัวอักษร ย.....	33
4.4 การวางผังแผงแป้นอักขระไทยแบบเกษมณี.....	34
4.5 การทำงานของระบบสนทนาภาษาไทยอัตโนมัติ	37
5.1 แสดงหน้าจอการทำงานของ A.L.I.C.E Bot.....	45
5.2 แสดงหน้าจอการติดต่อกับ A.L.I.C.E Bot เวอร์ชันภาษาไทยผ่านเบราว์เซอร์	46