บทที่ 6

ข้อมูลภาษาไทยที่ใช้อ้างอิงในการวัดประสิทธิภาพ

ในบทนี้จะได้กล่าวถึงข้อมูลภาษาไทยอ้างอิงที่ใช้เป็นอินพุทในการวัดประสิทธิภาพโดยเลือกข้อมูลจากสองแหล่งที่แตกต่างกัน กล่าวคือข้อมูลภาษาไทยที่ที่อยู่ในรปแบบอิเล็คทรอนิคส์และสามารถหาได้ทั่ว ๆ ไปโดยไม่ได้มีการคัดเลือกกลั่นกรองแจกแจงประเภทแต่อย่างใด กับข้อมูลอีกแบบหนึ่งที่ได้รับการคัดเลือกจัดหมวดหมู่ไว้เป็นอย่างดีซึ่งผู้วิจัยได้รับความอนุเคราะห์จากห้องวิจัยลิงคส์ (LINKS) ที่ เนคเทค (NECTEC) นอกจากนี้แล้วบทนี้ยังจะได้กล่าวถึงพจนานุกรมอ้างอิงที่จะใช้ในการตรวจสอบความถูกต้องของคำที่ตัดออกมาได้จากโปรแกรมตัดคำแต่ละแบบ

6.1 ฐานข้อมูลภาษาไทย

ได้ทำการเลือกฐานข้อมูลภาษาไทยจากสองแหล่งที่แตกต่างกันมาใช้เป็นอินพุทสำหรับการทดลองวัดประสิทธิภาพเพื่อเปรียบเทียบผลการทดลองที่ได้

6.1.1 ฐานข้อมูลภาษาไทยทั่ว ๆ ไป

สำหรับฐานข้อมูลภาษาไทยอ้างอิงนี้ได้มาจากเอกสารในรูปแบบอิเล็คทรอนิคส์ที่มีการเผยแพร่ไว้เป็นข้อมูลทั่ว ๆ ไปที่สามารถใช้งานได้ โดยไม่ได้เฉพาะเจาะจงกลุ่มใดเป็นพิเศษ และไม่ได้ทำการคัดเลือกแจกแจงประเภทแต่อย่างไร ซึ่งได้จากเอกสารต่าง ๆ ตามที่ระบุใน web page ดังที่แสดงไว้ข้างล่างนี้

1 แผนพัฒนาเศรฐกิจและสังคมแห่งชาติฉบับที่ 8 [1]

2 รายชื่องานวิจัย[2]

3 โครงการเทคโนโลยีสารสนเทศตามพระราชดำริ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี[3]

4 เทคโนโลยีสารสนเทศให้ประโยชน์อะไรต่อการพัฒนาประเทศไทย[1]

5 พระบาทสมเด็จพระเจ้าอยู่หัวกับเทคโนโลยีสารสนเทศ[3]

6 กฎหมายแลกเปลี่ยนข้อมูลทางสื่ออิเล็คทรอนิคส์[1]

7 พระราชพิธีกาญจนาภิเษก ฉลองสิริราชสมบัติ ครบ ๕๐ ปี พระบาทสมเด็จพระเจ้าอยู่หัวภูมิพลอดุลยเดช[3]

8 พระราชประวัติโดยสังเขปของพระบาทสมเด็จพระปรมินทรมหาภูมิพลอดุลยเดชมหาราช[3]

9 พระราชดำรัส พระราชทานแก่นายอุทัย พิมพ์ใจชน ประธานสภาร่างรัฐธรรมนูญ[3]

10 พระราชดำรัส พระราชทานแก่ปวงชนชาวไทยในโอกาสวันขึ้นปีใหม่[3]

11 พระราชดำรัสพระราชทานแก่คณะบุคคลต่าง ๆ ที่เข้าเฝ้าฯในโอกาสวันเฉลิมพระชนมพรรษา พ.ศ. 2539[3]

12 พระราชดำรัสพระราชทานแก่คณะบุคคลต่าง ๆ ที่เข้าเฝ้าฯในโอกาสวันเฉลิมพระชนมพรรษา พ.ศ. 2538[3]

13 พระราชดำรัสพระราชทานแก่คณะบุคคลต่าง ๆ ที่เข้าเฝ้าฯในโอกาสวันเฉลิมพระชนมพรรษา พ.ศ. 2537[3]

14 พระราชดำรัสพระราชทานแก่คณะบุคคลต่าง ๆ ที่เข้าเฝ้าฯในโอกาสวันเฉลิมพระชนมพรรษา พ.ศ. 2536[3]

15 พระราชดำรัสพระราชทานแก่คณะบุคคลต่าง ๆ ที่เข้าเฝ้าฯในโอกาสวันเฉลิมพระชนมพรรษา พ.ศ. 2535[3]

16 พระราชดำรัส พระราชทานแก่ พลเอกสุจินดา คราประยูร และ พลตรีจำลอง ศรีเมือง วันพุธที่ ๒๐ พฤษภาคม พ.ศ. 2535[3]

ที่มา:

1: National Information Technology Commitee, ก.ท.ส.ท., NITC (www.nitc.go.th)

2: ศูนย์เอกสารประเทศไทย จุฬาลงกรณ์มหาวิทยาลัย

3 เครือข่ายกาญจนาภิเษก (http://kanchanapisek.or.th)

6.1.2 ฐานข้อมูลอ้างอิงภาษาไทยที่ได้กลั่นกรองรวบรวมไว้

ฐานข้อมูลภาษาไทยที่ได้กลั่นกรองรวบรวมไว้อย่างเป็นระเบียบคือฐานข้อมูลไทยดาต้าแบงค์ (Thai Data Bank) ซึ่งได้รับการสร้างและรวบรวมโดยศูนย์วิจัยและพัฒนาเอไอ (AI Research and Development Center) ที่สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี (KMITT) ผู้วิจัยขอขอบพระคุณ อ. วันทนีย์ และ ดร. สุรพันธ์จากห้องปฏิบัติการลิงคส์ (LINKS - Linguistics and Knowledge Science Laboratory) ที่ศูนย์เทคโนโลยีอิเล็คทรอนิคส์และคอมพิวเตอร์แห่งชาติ หรือ เนคเทค (NECTEC) ที่ได้ให้คำปรึกษาที่เป็นประโยชน์มากต่องานวิจัยนี้และได้แนะนำฐานข้อมูลภาษาไทยนี้ให้กับผู้ทำวิจัย

ฐานข้อมูลภาษาไทยนี้ได้ถูกรวบรวมไว้เพื่อวัตถุประสงค์สามอย่างดังต่อไปนี้

1 เพื่อสร้างฐานข้อมูลภาษาไทยขนาดใหญ่

2 เพื่อให้บริการฐานข้อมูลสำหรับงานเขียนภาษาไทยมาตรฐาน

3 เพื่อรวบรวมงานศึกษาทุกแขนงและข้อมูลทั่วไปที่เกี่ยวกับประเทศไทย

ฐานข้อมูลภาษาไทยนี้ได้ถูกรวบรวมไว้ในช่วงระหว่างปี 1991 จนถึง 1994 โดยสามารถจัดกลุ่มแขนงได้ดังตาราง 6.1 สำหรับขนาดของฐานข้อมูลภาษาไทยนี้จะเป็นดังตาราง 6.2

แหล่งที่มา	สาขา
Reference book	Linguistics
Dictionary	Computer
Encyclopedia	Humanities
Textbook	Social Studies
Journal	Politics
Magazine	Art & Culture
Newspaper	Agriculture
Short Story	Business
Novel	Sciences
Others	Applied Sciences

ตาราง 6.1 การแบ่งกลุ่มของฐานข้อมูลภาษาไทยใน Thai Data Bank

THAI DATA BANK	average	quantity
Reference book	8.27%	3.97MB
Computer Journal/Textbook	24.04%	11.54MB
Journal	23.60%	11.33MB
Textbook	19.23%	9.23MB
Newspaper	15.88%	7.62MB
Novel	4.54%	2.18MB
Short Story	2.75%	1.32MB
Others	1.69%	0.81MB

ตาราง 6.1 แสดงขนาดของฐานข้อมูลภาษาไทยที่ใช้ในงานวิจัยนี้

6.2 พจนานุกรมอ้างอิง

พจนานุกรมอ้างอิงที่ใช้ในการตรวจสอบความถูกต้องของคำที่ได้จากการตัดคำด้วยโปรแกรมที่นำมาวัดประสิทธิภาพ จะได้ใช้พจนานุกรมฉบับราชบัณฑิตยสถานซึ่งรวบรวมโดย Mr. Doug Couper[1] มีขนาด 18,057 คำ แต่เนื่องจากได้พบว่าในพจนานุกรมชุดนี้มีคำที่มีความผิดพลาดปะปนอยู่จำนวนหนึ่งจึงได้ทำการลบคำเหล่านั้นออกไป เหลือที่ใช้งานจริง 17,859 คำ สำหรับพจนานุกรมที่ใช้โดยโปรแกรมตัดคำภาษาไทยในการตัดคำจะใช้พจนานุกรมที่ได้รวบรวมโดยผู้พัฒนาโปรแกรมนั้น ๆ เองถ้ามีมาให้เช่นพจนานุกรมของโปรแกรมตัดคำภาษาไทยแบบย้อนกลับจะมีขนาด 9,851 คำ เป็นต้น ตาราง 6.2 รวบรวมพจนานุกรมที่ใช้สำหรับโปรแกรมตัดคำแบบต่าง ๆ

โปรแกรมตัดคำ	จำนวนคำ	ที่มา
แบบเทียบคำสั้นสุด	17,859	Mr. Doug Couper
แบบเทียบคำยาวสุด	17,859	Mr. Doug Couper
แบบย้อนกลับ	9,851	สัมพันธ์ ระรื่นรมย์
แบบใช้ความถี่ของคำ	3,955	ดึงมาจากฐานข้อมูลภาษาไทยในหัวข้อ 6.1.1
แบบเลือกคำน้อยสุด	17,859	Mr. Doug Couper
แบบใช้พจนานุกรมคำกำกวม	21,990	Mr. Doug Couper ลบคำที่ทำให้เกิดความผิดพลาดมาก เช่น "การก" และเพิ่มคำศัพท์เพื่อแก้ปัญหาคำกำกวม

ตาราง 6.2 พจนานุกรมที่ใช้สำหรับโปรแกรมตัดคำแบบต่าง ๆ

หลังจากได้มีการพิจารณาฐานข้อมูลอ้างอิงภาษาไทยที่จะนำมาใช้ในการวิจัยแล้วในบทนี้ จะได้กล่าวถึงการออกแบบวิธีและขั้นตอนสำหรับการวัดประสิทธิภาพในบทต่อไป

ที่มา:

1 Southeast Asian Software Research Center, Bangkok 246-9311 (-28), Ext. 1617 doug@nwg.nectec.or.th
http://seasrc.th.net http://seasrc.th.net/sealang --> SEALANG Web site