บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

บทนี้จะกล่าวถึงเนื้อหา 2 ส่วน คือ ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 นิยามนิพจน์ระบุนาม

การสกัดนิพจน์ระบุนามหรือเนมเอนทิตี(Named Entity) เป็นคำนามที่เป็นชื่อเฉพาะในการเรียกชื่อบุคคล ชื่อกลุ่มบุคคล ชื่อองค์กร ชื่อสถานที่ วันเวลา และปริมาณ สำหรับหลักภาษาไทยได้กำหนดคำนิยามของชื่อเฉพาะ ไว้ดังนี้

พระยาอุปกิตศิลปสาร (2522 :71-72) ได้กล่าวถึงคำนามที่เป็นชื่อเฉพาะในเนื้อหาของคำนามประเภทวิสามานยนาม สามารถสรุปได้ดังนี้

วิสามานยนาม คำนามที่เป็นชื่อเฉพาะที่สมมติตั้งขึ้นสำหรับเรียก คน สัตว์ และสิ่งของบางอย่าง เพื่อให้รู้ชัดว่า คนนี้ สัตว์ตัวนี้ ของสิ่งนี้ ฯลฯ เช่น ตัวอย่าง ชื่อคน-สอน สิน ฯลฯ ชื่อบรรดาศักดิ์-ญาณภิรมย์ อุดมจินดา ฯลฯ ชื่อสกุล-มาลากุล ณ สงขลาฯ ชื่อสัตว์ เช่น ช้าง-มงคล พัง แป้น ฯลฯ ชื่อเมือง-นนทบุรี ราชบุรี ฯลฯ ชื่อสิ่งของ เช่น เรือ-เสือ ทะยานชน เสือคำรนสินธุ์ และวัน-อาทิตย์ จันทร์ ฯลฯ วิสามานยนามนี้ต้องเป็นคำใช้เป็นชื่อตั้งขึ้นเรียกคนคนเดียว สัตว์ตัวเดียว และของสิ่งเดียว ถึงจะเป็นชื่อของหมู่คณะ ก็ต้องเป็นหมู่เดียว คณะเดียว เช่น ‘ชาติ-ไทย’ หมายความว่าไทยชาติเดียว และ ’สโมสร-สามัคยาจารยสมาคม’ ก็หมายความถึงสโมสรนั้นแห่งเดียว ถึงแม้จะเผอิญมีชื่อซ้ำกันบ้าง ก็หมายความเฉพาะคน เฉพาะสิ่ง ไม่ทั่วถึงกันได้ ถ้าอยู่ใกล้กันก็ต้องเติมสร้อยหรือบอกเครื่องหมายท้ายชื่อให้สังเกตได้ว่า คนนี้ สิ่งนี้ เช่น นาย-แดง(เล็ก) นาย-แดง(ใหญ่) วัดสามจีนใต้ วันสามจีนเหนือ เป็นต้น

นพวรรณ พันธุเมธา(2527:4-5) ได้กล่าวถึงคำนามที่เป็นชื่อเฉพาะในเนื้อหาของคำนามประเภทคำนามวิสามัญ สามารถสรุปได้ดังนี้

คำนามวิสามัญ คือคำที่หมายถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะ เช่น บังอร กาญจนบุรี รักเร่ ตูบ

คำนามวิสามัญนั้นหมายถึงสิ่งใดสิ่งหนึ่งโดยเฉพาะอยู่แล้ว จึงมักไม่ต้องมีคำขยายช่วยจำกัดความหมาย นอกจากว่าคำนามวิสามัญคำใดใช้เรียกชื่อมากกว่า 2 สิ่ง จึงจำเป็นจะต้องหาคำอื่นมาจำกัดความหมายอีกทีหนึ่ง

ตัวอย่าง 1.) แมว รักเจ้าของ

2.) แมว ตัวนั้นรักเจ้าของ

3.) แต้ม รักเจ้าของ

4.) แต้ม ตัวนั้นรักเจ้าของ

แต้ม เป็นคำนามวิสามัญ กลุ่มคำ ตัวนั้น ซึ่งใช้จำกัดความหมายของคำนามปรากฏอยู่ในประโยคที่ 2 และประโยคที่ 4 ประโยคที่ 4 จะมีที่ใช้น้อยกว่าประโยคที่ 2 มาก เพราะคำว่า แต้ม มีความหมายเฉพาะอยู่แล้ว ไม่ควรจะต้องมีคำ ตัวนั้น ช่วยจำกัดความหมายลงไปอีก ประโยคที่ 4 จะทำให้เข้าใจว่า มีแมวที่ชื่อแต้ม หลายตัว ผู้พูดจึงต้องระบุว่าตัวไหน

กำชัย ทองหล่อ(2550:197) ได้กล่าวถึงคำนามที่เป็นชื่อเฉพาะในเนื้อหาของคำนามประเภทคำนามวิสามานยนาม สามารถสรุปได้ดังนี้

วิสามานยนาม คือคำนามที่เป็นชื่อเฉพาะของ คน สัตว์ สถานที่ และสิ่งของ เป็นชื่อที่บัญญัติขึ้นสำหรับใช้เรียกชี้เฉพาะลงไปว่า เป็นใครหรืออะไร

ตัวอย่าง 1.) สมศรี เป็นหญิงสาวที่เด่นในสังคม

2.) นาย แดง และนาย ดำ เป็นพี่น้อง

3.) เอราวัณ เป็นช้างอยู่ในสวรรค์ชั้น ดาวดึงส์

4.) เขาไปสมัครงานที่ กระทรวงศึกษาธิการ

5.) สามก๊ก เป็นหนังสือพงศาวดารจีน

2.1.2 ประเภทของนิพจน์ระบุนาม

1.) ประเภทชื่อบุคคล หมายถึงคำนามที่เป็นชื่อเฉพาะในการเรียกชื่อบุคคล เช่น นายอมรเทพ พวงไธสง นายอภิสิทธิ์ เวชชาชีวะ เป็นต้น

2.) ประเภทชื่อองค์กร หมายถึงคำนามที่เป็นชื่อเฉพาะในการเรียกชื่อกลุ่มบุคคล ชื่อองค์กร เช่น ธนาคารพัฒนาวิสาหกิจขนาดกลางและขนาดย่อมแห่งประเทศไทย กระทรวงกลาโหม เป็นต้น

3.) ประเภทชื่อสถานที่ หมายถึงคำนามที่เป็นชื่อเฉพาะในการเรียกชื่อสถานที่ ทั้งที่แบ่งตามภูมิประเทศและแบ่งตามการปกครอง เช่น ประเทศไทย , จังหวัดนครราชสีมา เป็นต้น

4.) ประเภทวันที่ หมายถึงกลุ่มคำที่บ่งบอกตามปฏิทิน เช่น 14 กุมภาพันธ์ 2553 , วันสงกรานต์ เป็นต้น ยกเว้นวันในสัปดาห์ที่ไม่สามารถระบุในปฏิทินได้ เช่น วันจันทร์ วันอังคาร เป็นต้น

5.) ประเภทเวลา หมายถึงกลุ่มคำที่บ่งบอกหน่วยของเวลา เช่น 19:00 น. เป็นต้น

6.) ประเภทปริมาณ หมายถึงกลุ่มคำที่บอกจำนวน เช่น ประมาณ 10 ชิ้น , 19 ชิ้น โดยประมาณ

2.1.3 **การเรียนรู้ของเครื่อง** (Machine Learning)

SWATH (Smart Word Analysis for Thai)

เป็นโปรแกรมการตัดคำภาษาไทยที่สามารถเลือกวิธี การตัดคำได้สองวิธี คือการตัดคำแบบเลือกคำที่ยาวที่สุด (longest matching) และการตัดคำโดยเลือกแบบเหมือนมากที่สุด (maximal matching algorithms) ซึ่งนอกเหนือจากการใช้งานได้ดีกับข้อความที่เป็น text ธรรมดา โปรแกรมยังสามารถรองรับไฟล์ในรูปแบบต่างๆได้แก่ html, rtf
(ที่มา: http://www.hlt.nectec.or.th/products/swath.php)

การตัดและกำกับชนิดของคำโปรแกรม SWATH มีรายละเอียดดังตารางที่ 2.1 ดังนี้

ตารางที่ 2.1
การกำกับชนิดของคำโปรแกรม SWATH

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
1	NPPR	คำนามประเภทวิสามานยนาม	วินโดวส์ 95 , โค้ก, พระอาทิตย์
2	NCNM	คำนามแสดงจำนวนไม่ใช่เชิงปริมาณ	หนึ่ง, สอง, สาม, 1,2,3
3	NONM	คำนามใช้บอกลำดับ	ที่หนึ่ง, ที่สอง, ที่1, ที่2

ตารางที่ 2.1 (ต่อ)

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
4	NLBL	คำนามใช้ป้ายกำกับรายการ	1, 2, 3, 4, ก, ข, a, b
5	NCMN	คำนามที่ใช้เรียกชื่อทั่วไป	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	คำนามแสดงตำแหน่งของบุคคล	ดร., พลเอก
7	PPRS	คำสรรพนาม	คุณ, เขา, ฉัน
8	PDMN	คำสรรพนามชี้เฉพาะ	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	คำสรรพนามที่เป็นคำถาม	ใคร, อะไร, อย่างไร
10	PREL	คำสรรพนามที่ใช้เป็นบทเชื่อมประโยค	ที่, ซึ่ง, อัน, ผู้
11	VACT	คำกริยาแสดงถึงการกระทำของตน	ทำงาน, ร้องเพลง, กิน
12	VSTA	คำกริยาที่อ้างถึงการบอกกล่าว	เห็น, รู้, คือ
13	VATT	คำกริยาวิเศษณ์	อ้วน, ดี, สาย
14	XVBM	คำช่วยกริยาก่อนหน้าคำปฏิเสธ "ไม่"	ฝนเกิดไม่ตก
15	XVAM	คำช่วยกริยาหลังคำปฏิเสธ "ไม่"	เขาไม่ค่อยมาที่นี่
16	XVMM	คำช่วยก่อนหรือหลังกริยาคำปฏิเสธ" ไม่"	เธอ(ไม่)ควรไปพบเขา หรือ เรา ควร(ไม่)พูดเลยวันนี้
17	XVBB	คำช่วยก่อนคำกริยาอยู่ตำแหน่งเริ่มต้นของประโยค	กรุณา, จง, เชิญ, อย่า, ห้าม
18	XVAE	คำช่วยที่ใช้ตามคำกริยา	ยกมือขึ้น, เด็กกินไปเล่นไป
19	DDAN	คำบ่งชี้ใช้โดยทันทีหลังคำนาม	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	คำบ่งชี้ตามคำนามหรือลักษณนาม	นี้, นั้น, โน้น, นู้น
21	DDBQ	คำบ่งชี้ใช้ระหว่างคำนามและคำลักษณนามหรือก่อนคำแสดงปริมาณ	ทั้ง, อีก, เพียง
22	DDAQ	คำบ่งชี้ใช้หลังคำแสดงปริมาณ	พอดี, ถ้วน
23	DIAC	คำไม่เฉพาะเจาะจงใช้หลังคำนามหรือนอกเหนือระหว่างคำลักษณะนาม	ไหน, อื่น, ต่างๆ

ตารางที่ 2.1 (ต่อ)

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
24	DIBQ	คำไม่เฉพาะเจาะจงใช้ระหว่างคำนามและคำลักษณะนามหรือก่อนคำแสดงปริมาณ	บาง, ประมาณ, เกือบ
25	DIAQ	คำไม่เฉพาะเจาะจงใช้ตามคำแสดงปริมาณ	กว่า, เศษ
26	DCNM	คำบ่งชี้ของตัวเลขแสดงจำนวนใช้ในการแสดงคำปริมาณ	หนึ่งคน, เสือ 2 ตัว
27	DONM	คำบ่งชี้ของเลขลำดับใช้ในการแสดงคำปริมาณ	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	คำวิเศษณ์แบบปกติ	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	คำวิเศษณ์แบบซ้ำโดยรวม "ๆ"	เร็วๆ, เสมอๆ, ช้าๆ
30	ADVP	เพิ่มคำนำหน้าคำวิเศษณ์	โดยเร็ว
31	ADVS	คำวิเศษณ์แสดงทัศนคติของผู้พูดหรือการประเมินสิ่งที่กล่าวในส่วนที่เหลือของประโยค	โดยปกติ, ธรรมดา
32	CNIT	หน่วยบอกลักษณะนาม	ตัว, คน, เล่ม
33	CLTV	ลักษณะนามที่ใช้ในการแสดงชุด, กลุ่ม, ระดับหรือประเภทของสัตว์หรือบุคคล	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน
34	CMTR	ลักษณะนามแสดงหน่วยวัดปริมาณ	กิโลกรัม, แก้ว, ชั่วโมง
35	CFQC	ลักษณะนามแสดงความถี่	ครั้ง, เที่ยว
36	CVBL	ส่วนขยายเพื่อแสดงหน่วยของคำนาม	ม้วน, มัด
37	JCRG	คำสันธานเชื่อมประโยคระดับเดียวกัน	และ, หรือ, แต่
38	JCMP	คำสันธานเชื่อมการเปรียบเทียบ	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	คำสันธานเชื่อมประโยคย่อย	เพราะว่า, เนื่องจาก, ที่, แม้ว่า

ตารางที่ 2.1 (ต่อ)

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
40	RPRE	คำบุพบท	จาก, ละ, ของ, ใต้, บน
41	INT	คำอุทาน	โอ้ย, โอ้, เออ, เอ๋, อ๋อ
42	FIXN	คำนำหน้าคำนามใช้บอกจุดเริ่มต้นคำกริยา มี 2 คำ "การ", "ความ"	การทำงาน, ความสนุกสนาน
43	FIXV	คำนำหน้ากริยาวิเศษณ์ มี 1 คำ "อย่าง"	อย่างเร็ว
44	EAFF	คำที่บันทึกอยู่ในท้ายประโยคเพื่อแสดงอารมณ์ของคำพูด	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า
45	EITT	คำที่บันทึกอยู่ในท้ายประโยคเพื่อแสดงอารมร์ของคำถาม	หรือ, เหรอ, ไหม, มั้ย
46	NEG	คำปฏิเสธ	ไม่, มิได้, ไม่ได้, มิ
47	PUNC	อักขระพิเศษ	(,), ", ,, ;

ที่มา : http://www.hlt.nectec.or.th//orchid

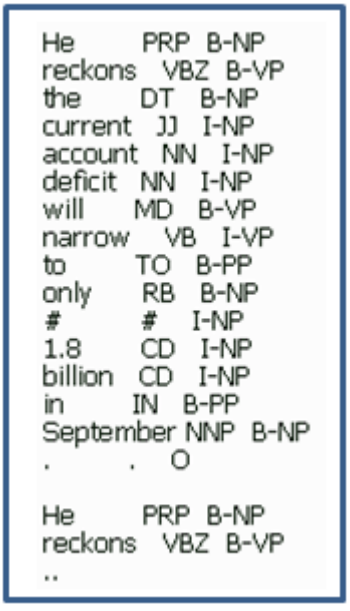
Conditional Random Fields (CRFs)

เป็นเทคนิคการเรียนรู้วิธีหนึ่ง โดยจะมีโปรแกรม CRF++ สำหรับใช้ฝึก (training) และทดสอบ (test) ระบบ เทคนิค CRFs จะใช้ในการตัดคำหรือติดลาเบล (label) ลำดับข้อมูล ถูกออกแบบมาเพื่อวัตถุประสงค์ทั่วไปและสามารถนำไปประยุกต์ใช้กับงานทางด้านการประมวลผลภาษาธรรมชาติ เช่น การรู้จำนิพจน์ระบุนาม (Named Entity Recognition), การสกัดสารสนเทศ (Information Extraction) โปรแกรม CRF++ สามารถกำหนดชุดคุณสมบัติเองได้ มีความเร็วในการฝึกฝน(training)ระบบ อีกทั้งยังใช้หน่วยความจำน้อย และสามารถแสดงเอาพุต (Output) ความน่าจะเป็นของขอบเขตทั้งหมดของแคนดิเดต(Candidates)ได้

รูปแบบการฝึกฝนและทดสอบไฟล์ (Training and Test file formats)

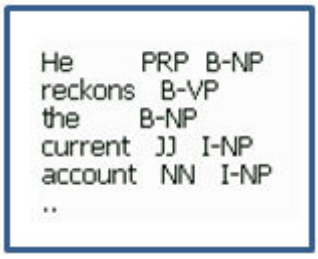
การฝึกฝนและทดสอบไฟล์ต้องประกอบไปด้วยคำจำนวนมาก โดยแต่ละคำต้องแทนในหนึ่งบรรทัดและใช้ช่องไฟหรือแท็บในการแบ่งแต่ละคอลัมน์ การเรียงลำดับของแต่ละคำมาจากข้อความในประโยค

ภาพที่ 2.1
ตัวอย่างไฟล์สำหรับการฝึกและทดสอบ



จากตัวอย่างไฟล์สำหรับฝึกและทดสอบใน ภาพที่ 2.1 คอลัมน์ที่ 1 เป็นคำที่นำมาจากข้อความของประโยค คอลัมน์ที่ 2 เป็นการกำกับชนิดของคำหรือคุณสมบัติที่ใช้สำหรับการสกัดหรือระบุนิพจน์ระบุนาม และคอลัมน์ที่ 3 เป็นการบอกเงื่อนไขคำนิพจน์ระบุนาม กำหนดให้อักขระ B เป็นการบอกจุดเริ่มต้นของคำที่ต้องการสกัด อักขระ I เป็นการบอกจุดต่อเนื่องที่ต้องการสกัด และ O เป็นการบอกจุดของคำที่ไม่ต้องการสกัด

ภาพที่ 2.2
ตัวอย่างไฟล์การฝึกและทดสอบที่ไม่ถูกต้อง

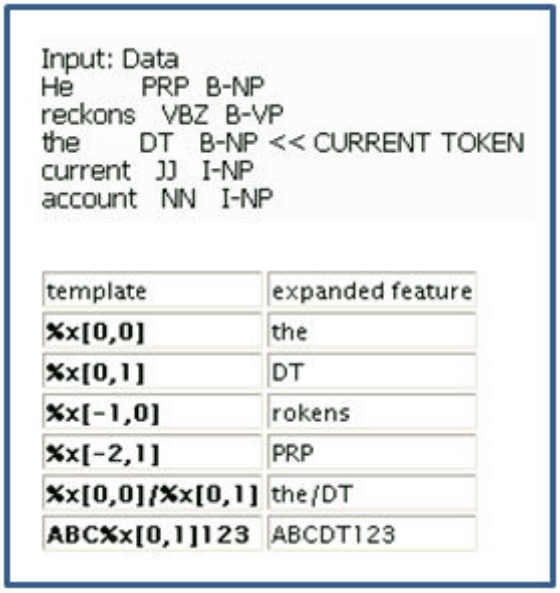


ตัวอย่างไฟล์ในภาพที่ 2.2 เป็นการสร้างไฟล์สำหรับการฝึกและทดสอบที่ไม่ถูกต้องตามรูปแบบมาตรฐาน เนื่องจากบรรทัดที่ 2 และ 3 มีแค่ 2 คอลัมน์ ไม่มีการกำกับชนิดของคำหรือคุณสมบัติที่ใช้สำหรับการสกัดข้อมูล

การจัดเตรียมคุณสมบัติของต้นแบบ (Preparing feature templates) ในแต่ละบรรทัดจะแสดงหนึ่งต้นแบบ มี macro %x[row , col] ที่ใช้ในการระบุคำของข้อมูลที่ถูกป้อนเข้ามา โดยกำหนดให้ row ใช้ในการระบุตำแหน่งของคำปัจจุบันที่โฟกัสและ col ใช้เลือกตำแหน่งคอลัมน์ของแถวที่โฟกัส

ภาพที่ 2.3

ตัวอย่างการค้นหา template



Input: Data
He PRP B-NP
reckons VBZ B-VP
the DT B-NP << CURRENT TOKEN
current JJ I-NP
account NN I-NP

template	expanded feature
%x[0,0]	the
%x[0,1]	DT
%x[-1,0]	rokens
%x[-2,1]	PRP
%x[0,0]/%x[0,1]	the/DT
ABC%x[0,1]123	ABCDT123

template จะมีอยู่ 2 ประเภทคือคำที่ขึ้นต้นด้วยตัวอักษร U หมายถึง Unigram Template เช่น “Uo1:%[0 ,1]” โปรแกรม CRF++ จะสร้างชุดคุณสมบัติของฟังก์ชันให้แบบอัตโนมัติโดยที่จำนวนคุณสมบัติของฟังก์ชันที่สร้างจะมีจำนวนเท่ากับ L * N และคำที่ขึ้นต้นด้วยตัวอักษร B หมายถึง Bigram Template โปรแกรม CRF++ จะทำการสร้างแบบอัตโนมัติโดยจะมีการรวมกันของ output ของคำปัจจุบันและ output ของคำที่ผ่านมาซึ่งคุณสมบัติจะต้องไม่ซ้ำกันเลยมีจำนวนเท่ากับ (L * L * N)

กำหนดให้

L คือจำนวนของ output tag

N คือจำนวนของคำที่ไม่ซ้ำกันจาก template สร้างให้

ภาพที่ 2.4

ตัวอย่างโปรแกรม CRF++ สร้างชุดอัตโนมัติของคุณสมบัติฟังก์ชัน

```
func1 = if (output = B-NP and feature="U01:DT") return 1 else return 0
func2 = if (output = I-NP and feature="U01:DT") return 1 else return 0
func3 = if (output = O and feature="U01:DT") return 1  else return 0
....
funcXX = if (output = B-NP and feature="U01:NN") return 1  else return 0
funcXY = if (output = O and feature="U01:NN") return 1  else return 0
...
```

ภาพที่ 2.5

ตัวอย่างไฟล์ Template

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]q
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

# Bigram
B
```

การฝึกระบบ(Training)

ภาพที่ 2.6

คำสั่งฝึกระบบ crf_learn

```
crf_learn template_file train_file model_file
```

การฝึกระบบจะมีรูปแบบคำสั่งมาตรฐานสำหรับการฝึกระบบดังนี้

crf_learn เป็นโปรแกรมสำหรับการฝึก

template_file และ train_file ผู้ใช้จะต้องเป็นคนจัดเตรียมเอง

model_file โปรแกรม crf_learn จะทำการสร้างให้

ภาพที่ 2.7

ตัวอย่าง Output คำสั่ง crf_learn

```
CRF++: Yet Another CRF Tool Kit
Copyright(C) 2005 Taku Kudo, All rights reserved.

reading training data: 100.. 200.. 300.. 400.. 500.. 600.. 700.. 800..
Done! 1.94 s

Number of sentences: 823
Number of features:  1075862
Number of thread(s): 1
Freq:                1
eta:                 0.00010
C:                   1.00000
shrinking size:      20
Algorithm:           CRF

iter=0 terr=0.99103 serr=1.00000 obj=54318.36623 diff=1.00000
iter=1 terr=0.35260 serr=0.98177 obj=44996.53537 diff=0.17161
iter=2 terr=0.35260 serr=0.98177 obj=21032.70195 diff=0.53257
iter=3 terr=0.23879 serr=0.94532 obj=13642.32067 diff=0.35138
iter=4 terr=0.15324 serr=0.88700 obj=8985.70071 diff=0.34134
iter=5 terr=0.11605 serr=0.80680 obj=7118.89846 diff=0.20775
iter=6 terr=0.09305 serr=0.72175 obj=5531.31015 diff=0.22301
iter=7 terr=0.08132 serr=0.68408 obj=4618.24644 diff=0.16507
iter=8 terr=0.06228 serr=0.59174 obj=3742.93171 diff=0.18953
```

ภาพที่ 2.7 Output ที่ได้จากคำสั่ง **crf_learn** มีรายละเอียดดังนี้

iter คือจำนวนของรอบที่โปรเซส

terr คือจำนวนของความผิดพลาดของแท็ก

serr คือจำนวนของความผิดพลาดของประโยค

obj คือค่าของอ๊อบเจ็กปัจจุบันเมื่อค่านี้เข้าใกล้จุดที่กำหนด

diff คือความแตกต่างจากค่าอ๊อบเจ็กที่ผ่านมา

การทดสอบ(Testing)

ภาพที่ 2.8

คำสั่งทดสอบระบบ crf_test

```
crf_test -m model_file test_file
```

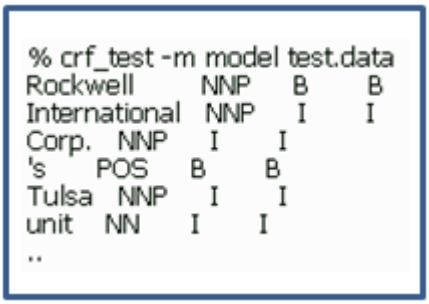
การทดสอบระบบจะมีรูปแบบคำสั่งมาตรฐานสำหรับการทดสอบดังนี้

crf_test เป็นโปรแกรมสำหรับการทดสอบ

model_file เป็นโมเดลที่เกิดจาก crf_learn ทำการสร้างให้

ในการทดสอบไม่ต้องมี template_file เนื่องจากใน model_file มีข้อมูลของ template อยู่แล้ว สำหรับข้อมูลที่จะใช้ในการทดสอบรูปแบบของไฟล์จะต้องเขียนเหมือนกับรูปแบบของ training_file

ภาพที่ 2.9
ตัวอย่าง Output คำสั่ง crf_test



ในภาพที่ 2.9 ผลจากการทดสอบสามารถประเมินค่าความถูกต้องได้จากความแตกต่างระหว่าง 2 คอลัมน์ โดยเปรียบเทียบระหว่างคอลัมน์ 3 และคอลัมน์ 4

Thai Lexeme Analyser (Tlexs)

ทีเล็กส์เป็นโปรแกรมแบ่งคำภาษาไทย ซึ่งพัฒนาโดยใช้เทคนิคการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการของ Conditional Random Field (CRF) ในการเรียนรู้และใช้คลังข้อมูลของ BEST2009 ขนาด 5 ล้านคำ ในการฝึกฝนโปรแกรมทีเล็กส์ (ที่มา: http://www.hlt.nectec.or.th)

Thai Lexeme Tokenizer (LexTo)

เล็กซ์โตเป็นโปรแกรมตัดคำสำหรับข้อความภาษาไทย จะรวมวิธีตัดคำแบบ Conditional Random Field กับการตัดคำแบบใช้พจนานุกรมร่วมกันทำงานเพื่อใช้แก้ปัญหาการตัดคำที่กำกวม

เหมืองข้อความ (Text Mining)

เป็นกระบวนการที่กระทำกับข้อความ(โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง หลักคณิตศาสตร์ หลักการประมวลเอกสาร (Document Processing) หลักการประมวลผลข้อความ (Text Processing) และหลักการประมวลผลภาษาธรรมชาติ (Natural Language Processing)

ความรู้ที่ได้จากการทำเหมืองข้อความมีหลายรูปแบบได้แก่

การสรุปเอกสารข้อความ (Document Summarization) เป็นการลดความซับซ้อนและขนาดของเอกสารข้อความโดยไม่ทำให้ความหมายหรือสาระสำคัญของข้อมูลเอกสารสูญเสียไป

การแบ่งประเภทเอกสารข้อความ (Document classification) จัดแบ่งประเภทของกลุ่มเอกสารข้อความออกเป็นคลาส โดยการใช้ชุดข้อมูลตัวอย่างของเอกสารข้อความที่เรียกว่า Training Set สำหรับสร้าง Classifier Model และทดสอบ Classifier Model ด้วย Test Set ขั้นตอนวิธีได้แก่ Supervised Learning Neural Network, C4.5 Decision Tree

การแบ่งกลุ่มเอกสารข้อความ (Document clustering) จัดแบ่งเอกสารข้อความออกเป็นกลุ่ม โดยใช้การวัดความคล้ายคลึงและความแตกต่างของคุณลักษณะของเอกสารข้อความ เพื่อนำไปใช้ประโยชน์ในด้านการข่าว ข้อมูลเอกสารจะถูกแปลงให้เป็นชุดข้อมูลตัวเลขโดยวิธีการ DFxIDF (Vector Space Model) จากนั้นถึงใช้ขั้นตอนวิธีการแบ่งกลุ่มข้อมูล ได้แก่ K-Mean, Unsupervised Learning Neural Networks, Hierarchical Clustering

(ที่มา : th.wikipedia.org/wiki/การทำเหมืองข้อความ)

2.2 งานวิจัยที่เกี่ยวข้อง

บุญเสริม กิจศิริกุล , ไพศาล เจริญพรสวัสดิ์ และ สุรพันธ์ เมฆนาวิน 1999:

งานระบุเนมด์เอนทิทีในอดีตที่ผ่านมาจะใช้การทำด้วยมือ ผู้วิจัยจึงได้นำเสนออัลกอริธึมสำหรับการเรียนรู้เพื่อช่วยพัฒนาการทำงานแบบอัตโนมัติ โดยนำเสนอการเปรียบเทียบอัลกอริธึมสำหรับการเรียนรู้สองอัลกอริธึมระหว่าง อัลกอริธึม Winnow กับ อัลกอริธึม RIPPER ผลการทดลองแสดงให้เห็นประสิทธิภาพของอัลกอริธึม Winnow ดีกว่า อัลกอริธึม RIPPER

อมรทิพย์ กวินปณิธาน 2546:

งานวิจัยมีวัตถุประสงค์เพื่อศึกษาบริบทที่สามารถเป็นบริบทบ่งบอกการปรากฏของชื่อเฉพาะภาษาไทยสำหรับนำไปใช้ในการพัฒนางานด้านภาษาศาสตร์คอมพิวเตอร์ผลการศึกษาชื่อเฉพาะทั้งหมด 2,547 ชื่อ พบว่าชื่อเฉพาะที่ปรากฏบริบทบ่งบอกคิดเป็น 63% ชื่อเฉพาะที่ไม่ปรากฏบริบทบ่งบอกคิดเป็น 37% บริบทที่สามารถบ่งบอกชื่อเฉพาะภาษาไทยได้แก่ (1) คำที่ปรากฏติดกับชื่อเฉพาะหรือคำบ่งบอก (2) การเว้นวรรค และ (3) ตัวบ่งบอกระดับปริจเฉท ชื่อที่ไม่ปรากกฏบริบทบ่งบอกพบว่า เป็นชื่อเฉพาะที่มีโครงสร้างของรูปภาษาในลักษณะทับศัพท์ภาษาต่างประเทศ หรือเป็นชื่อเฉพาะที่ผู้ใช้ภาษามีภูมิความรู้ร่วมกัน จึงเอื้อต่อการนำชื่อเฉพาะนั้นไปใช้โดยไม่ต้องมีตัวบ่งบอกในการระบุว่ารูปภาษานั้นเป็นชื่อเฉพาะ

หัชทัย ชาญเลขาและอัศนีย์ ก่อตระกูล 2546:

งานวิจัยนี้เสนอแนวทางการสกัดเนมด์เอนทิทีภาษาไทย โดยใช้แมกซิมัมเอนโทรปี โมเดลร่วมกับกฎและคลังคำศัพท์ รวมทั้งใช้สถิติของคำจากคลังเอกสารเพื่อหาตำแหน่ง ขอบเขต และประเภทของเนมด์เอนทิที การทดสอบประสิทธิภาพของระบบ พบว่ามีค่า F เท่ากับ 85.29% 82.67% และ 82.43% สำหรับชื่อบุคคล องค์กร และสถานที่ ตามลำดับ

Fuchun Peng, Fangrfang Feng และ Andrew McCallum 2004:

การตัดคำในภาษาจีนก็ประสบปัญหาเดียวกับการตัดคำในภาษาไทย เพราะลักษณะของภาษาจีนนั้นเป็นภาษาที่เขียนติดกันไปทั้งประโยคเช่นเดียวกับภาษาไทย จึงมีงานวิจัยที่ศึกษาเกี่ยวกับวิธีการต่าง ๆ ในการตัดคำของภาษาจีน สำหรับในงานของ Fuchun Peng, Fangfand Feng และ Andrew McCallum นั้น ได้ทำการสาธิตเอาเทคนิค Conditional Random Fields (CRFs) มาประยุกต์ใช้ในการตัดคำของภาษาจีน งานวิจัยทำการสร้างฐานความรู้ให้กับเครื่องโดยกำหนด Start Tag ให้กับตัวอักษรตัวแรก และกำหนด Non-start Tag ให้กับตัวอักษรตัวถัดไปในคำ นอกจากนี้มีการกำหนดหน้าที่ของคำให้กับแต่ละตัวอักษรด้วย จากผลงานวิจัยพบว่า CRFs สามารถช่วยเพิ่มประสิทธิภาพในการตัดคำในภาษาจีน

ภาพที่ 2.10
สัญลักษณ์ที่ใช้ในการกำหนดหน้าที่ตัวอักษร

C_{-2}:	second previous character in lexicon
C_{-1}:	previous character in lexicon
C_1:	next character in lexicon
C_2:	second next character in lexicon
C_0C_1:	current and next character in lexicon
$C_{-1}C_0$:	current and previous character in lexicon
$C_{-2}C_{-1}$:	previous two characters in lexicon
$C_{-1}C_0C_1$:	previous, current, and next character in the lexicon

สุฤดี ฉัตรไตรมงคล 2548:

งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้แนวทางแบบลูกผสม (hybrid approach) โดยแนวทางดังกล่าวจะแบ่งออกเป็นสองส่วนคือส่วนที่เป็นระบบทางสถิติและส่วนที่เป็นระบบกฎ สำหรับส่วนของระบบทางสถิตินั้นจะใช้วิธีทางสถิติร่วมกับโลคอลแมกซ์อัลกอริทึมเพื่อคัดเลือกกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะออกมา ส่วนระบบกฎจะถูกเขียนขึ้นโดยอิงกับหลักฐานที่ได้จากบริบทภายใน เช่น คำนำหน้าชื่อและบริบทข้างเคียง จากการทดสอบพบว่าระบบกฎที่สร้างขึ้นสามารถจำแนกประเภทของชื่อเฉพาะโดยให้อัตราการรู้จำ (ค่า F) สำหรับชื่อเฉพาะประเภทชื่อคน 69.15% ชื่อองค์กร 62.95% และชื่อสถานที่ 38.87% ตามลำดับ โดยมีค่าความแม่นยำและค่าความครบถ้วนสำหรับชื่อเฉพาะประเภทชื่อคน 54.00% และ 96.12% ชื่อองค์กร 47.60% และ 92.93% ชื่อสถานที่ 31.67% และ 50.32% ตามลำดับ

Nuttida Suwanno,Yoshimi Suzuki, Haruaki Yamazaki 2007:

งานวิจัยนี้นำเสนอวิธีการสกัดเนมด์เอนทิทีโดยใช้อัลกอริทึม support vector machine (SVMs) งานวิจัยจะเน้นที่เอนทิทีภาษาไทยประเภทบุคคล องค์กร และสถานที่ วิธีที่ใช้จะทำงานร่วมกันระหว่างวิธี rule-based และวิธี learning สำหรับการเลือกคำของประโยคจะใช้ขนาดหน้าต่างบวกลบ 2 ชุดของคุณสมบัติที่ใช้สำหรับการทดลองจะมี 4 ประเภทคือ word, Part of Speech(POS), Semantic Concept, Orthographic ผลลัพธ์ที่ได้แสดงให้เห็นว่ามีความถูกต้องในการสกัดเนมด์เอนทิทีภาษาไทยประเภทบุคคล 89% ประเภทสถานที่ 93% และประเภทองค์กร 76% และชุดคุณสมบัติที่ดีที่สุดสำหรับการทดลองนี้คือการรวมกันของ word , Semantic concept และ Orthographic features

Toru Hirano, Yoshihiro Matsuo, Genichiro Kikui

งานวิจัยนี้เสนอวิธีการจัดการเรียนรู้เพื่อค้นหาความหมายที่มีความเกี่ยวของความสัมพันธ์กันของเนมด์เอนทิทีที่อยู่ในต่างประโยค โดยจะใช้คุณสมบัติพื้นฐานของเนื้อหา (Contextual) กับ Centering Theory พร้อมกับการสร้างประโยค(syntactic)และพื้นฐานของคุณสัมบัติของคำ (word-base features) ซึ่งคุณสมบัตินี้มีการจัดการเป็นโครงสร้างต้นไม้ (tree structure) และมีการ fed ภายใน boosting – based classification algorithm จากผลการทดลองแสดงให้เห็นว่าให้ผลดีกว่าวิธีก่อนและยังเพิ่ม precision และ recall เป็น 4.4% และ 6.7 %

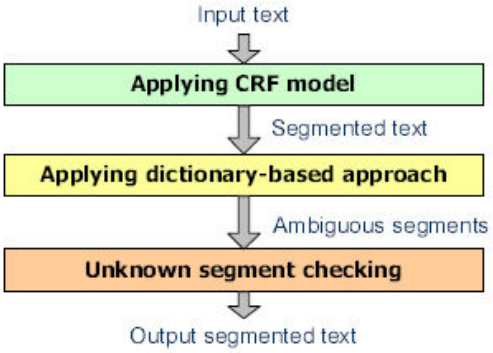
John Lafferty, Andrew McCallum และ Fernando Pereira

เป็นงานวิจัยที่นำเสนอเอาเทคนิค Conditional Random Fields (CRFs) มาช่วยในการแบ่งแยกข้อมูลที่อยู่ต่อเนื่องกัน โดยทำการเปรียบเทียบประสิทธิภาพกับวิธี Maximum Entropy , Markov Models (MEMMs) โดยข้อมูลที่นำมาทำการวิจัยถูกสร้างจาก Hidden Markov Models (HMMs) ซึ่งมี 2,000 ข้อมูลที่ใช้ในการเรียนรู้ (Train) และ 500 ข้อมูลที่ใช้ในการทดสอบผลจากงานวิจัยนี้คือ CRFs มีข้อผิดพลาด (error) น้อยกว่า MEMMs โดย CRFs มีค่าผิดพลาดเท่ากับ 4.6% ในขณะที่ MEMMs มีค่าผิดพลาดถึง 42%

ชูชาติ หไชยะศักดิ์, ศราวุธ คงยัง และชัยอนันต์ ดำรงรัตน์ (2008)

จากงานวิจัยสรุปออกมาว่า วิธี Conditional Random Field เป็นวิธีตัดคำแบบใช้เครื่องเรียนรู้ที่ดีที่สุด โดยงานวิจัยนี้ก็จะมุ่งประเด็นไปที่การรวมวิธีการตัดคำแบบ Conditional Random Field กับ การตัดคำแบบใช้พจนานุกรม เพื่อที่จะแก้ไปหาในส่วนของการตัดคำที่กำกวม โดยเรียกวิธีนี้ว่า LearnLexTo

ภาพที่ 2.11
ขั้นตอนการวิจัยของ LearnLexTo



LearnLexTo จะให้ CRF เป็นขั้นตอนหลักในการตัดคำ และให้วิธีที่ใช้พจนานุกรมเป็นตัวตรวจสอบข้อมูลที่ถูกตัดมาแล้วจาก CRF ดังภาพที่ 2.11

ในขั้นตอนแรกของการทดลองจะมีการประมาณค่าประสิทธิภาพของ CRF ถึงความเหมาะสมในการใช้จำนวนแกรม(Gram) จากการทดลองสรุปออกมาว่าค่า 11-Gram มีความเหมาะสมมากที่สุด และได้ค่า F-measure metric 85.7% ขั้นตอนถัดมาคือ การทดลองประสิทธิภาพของ LearnLexTo ซึ่งผลที่ได้คือ LearnLexTo มีค่า F-measure metric 87.67% ในขณะที่การตัดคำแบบใช้พจนานุกรมอย่างเดียวมีค่า F-measure metric เพียง 82.71%

Nutcha Tirasaroi, and Wirote Aroonmanakum 2009:

งานวิจัยนี้นำเสนอระบบการรู้จำนิพจน์ระบุนาม (Thai named entity recognition) โดยใช้ Conditional Random Fields (CRFs) โดยจะมีการเปรียบเทียบ input ที่ได้จากการ word-segmented กับ input syllable-segmented ผลลัพธ์จากการทดลองแสดงให้เห็นว่า syllable-segmented ดีกว่า word-segmented เล็กน้อย

Choochart Haruechaiyasak, Prapass Srichaivattana, Sarawoot Kongyoung and Chaianun Damrongrat

งานวิจัยนี้เสนอวิธีเลือกการตัดคำสำหรับการสกัด keywords ที่สำคัญจากประเภทคลังข้อความ โดยใช้อัลกอริธึมที่เรียกว่า Automatic Categorized Keyword Extraction (ACKE) อัลกอริธึมนี้จะใช้วิธีรูปแบบของเหมืองลำดับประกอบไปด้วย 2 ขั้นตอนหลักคือ กระบวนการสร้างรูปแบบความถี่ของ substring และกระบวนการ merging ความถี่ของ substring ภายใน keywords

Raymond J. Mooney and Un Yong Nahm

งานวิจัยนี้นำเสนอกรอบการทำงานสำหรับเหมืองข้อความที่เรียกว่า DISCOTEX (Discovery from Text EXtraction) ใช้การเรียนรู้ระบบสกัดสารสนเทศเพื่อแปลงข้อความในโครงสร้างข้อมูลที่เพิ่มเติมสำหรับความสัมพันธ์ที่สนใจ เวอร์ชันแรกของ DISCOTEX จะ integrateเข้ากับโมดูล IE โดยระบบการเรียนรู้ของ IE และกฎมาตรฐานของ induction โมดูล นอกจากนี้ กฎการสกัดฐานข้อมูลจากคลังข้อมูลของข้อความคือใช้การทำนายข้อมูลเพิ่มเติมจากเอกสารในอนาคต