

การวิเคราะห์เปรียบเทียบเครื่องมือแบ่งคำภาษาไทย และปัจจัยคุณลักษณะเฉพาะของอักขระ

โดย

ธิปกร ธนกุลวรภาส

ศราวุธ คงยัง

กฤษณ์ โกสวัสด์

หน่วยปฏิบัติการวิจัยวิทยาการมนุษยภาษา

ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

{krit.kos, sarawoot.kon, tipraporn.tha}@nectec.or.th

Outline

- ที่มาของ BEST และคลังข้อความ
- เครื่องมือแบ่งคำภาษาไทย
- ระบบที่เล็คซ์_256
- ผลการทดลอง
- สรุปผลการทดลองและข้อเสนอแนะ

ที่มาของ BEST และ คลังข้อความ

ผลการแข่งขัน BEST 2009

ระบบ	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความเหนียว
DHM	98.22	97.67	97.95
CUWS	98.10	97.58	97.84
KAM Thai	97.97	96.42	97.20
CPSK Cut	95.69	95.87	95.78
Tri-Gram	96.03	95.10	95.56
CRF-Technique	91.65	92.39	92.02

- คลังข้อความชุดฝึกฝน
ขนาด 5 ล้านคำ
- คลังข้อความชุดทดสอบ
ขนาด 6 แสนคำ
- คลังข้อความแบ่งเป็น 4 หมวด
1) สารานุกรมไทย 2) ข่าว
3) บทความวิชาการ 4) นิยาย

เครื่องมือแบ่งคำภาษาไทย

- ระบบจำลองผสมแบบแยกแยะ (DHM)
- ระบบควอร์ส (CUWS)
- ระบบคำไทย (KAM Thai)
- ระบบทีเล็กซ์_10 (TLexs_10)

ระบบจำลองผสมแบบแยกแยะ (DHM)

1. การสร้างแบบจำลองผสมแบบแยกแยะ

- ระดับคำ (Word-Level) กำกับชนิดของคำด้วย “W”

- ระดับกลุ่มอักขระ (Character-Level) กำกับอักขระด้วยสัญลักษณ์ 4 ตัว คือ

“B : เริ่มต้นคำ”

“I : อยู่ในคำ”

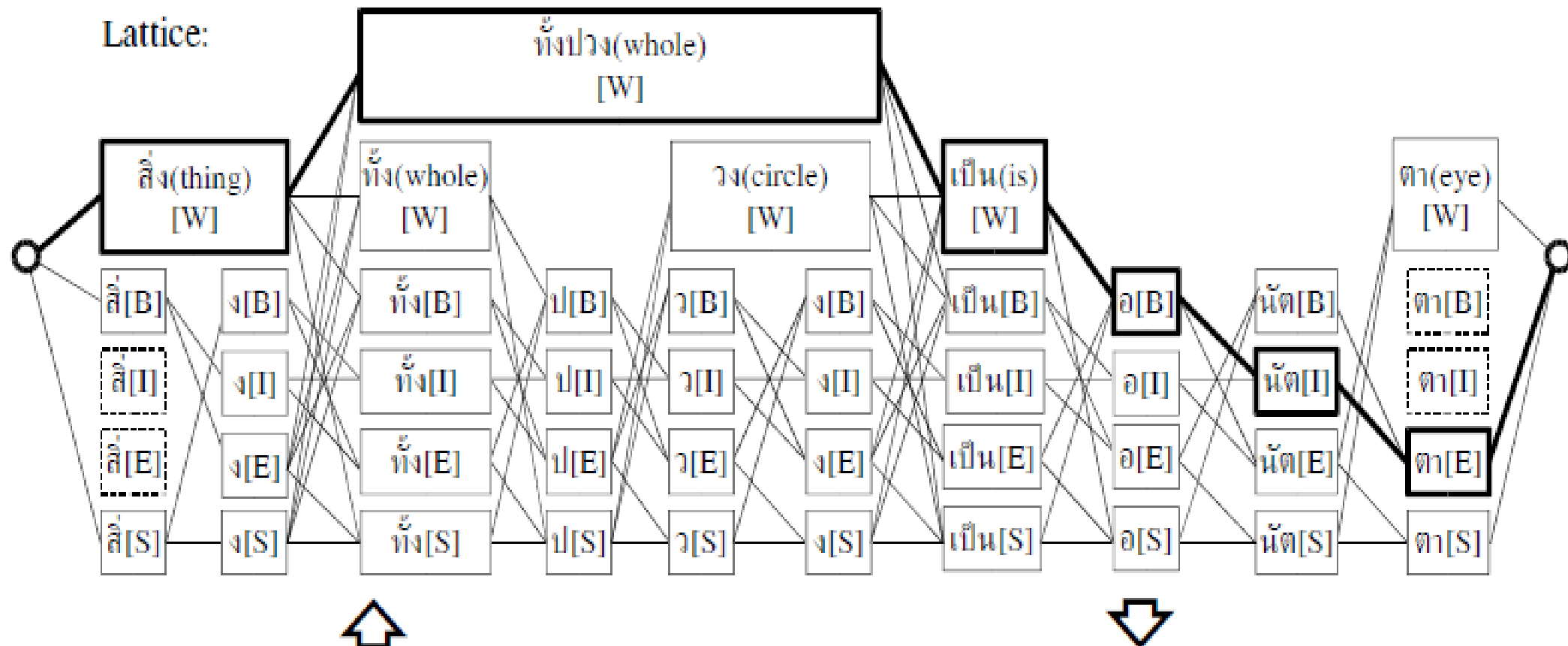
“E : อยู่ท้ายคำ”

“S : อักขระเดียว”

6

ระบบจำลองผสมแบบแยกแยะ (DHM) ต่อ

Lattice:



Input: สิ่งทั้งปวงเป็นอนัตตา
(Everything is no-self.)

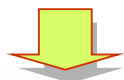
Output: สิ่ง|ทั้งปวง|เป็น|อนัตตา

ระบบจำลองผสมแบบแยกแยะ (DHM) ต่อ

2. ออนไลน์ลาร์จมาร์จิน เป็นการเรียนรู้แบบ
ออนไลน์ที่เรียกว่า “Margin Infused Relaxed
Algorithm” หรือ MIRA โดยระบบนี้ทำการ
สร้าง Lattice ขึ้น เพื่อสร้างเป็นแบบจำลองให้
ระบบสามารถเรียนรู้ได้

ระบบคิวส์ (CUWS)

ข้อความ
ภาษาไทย



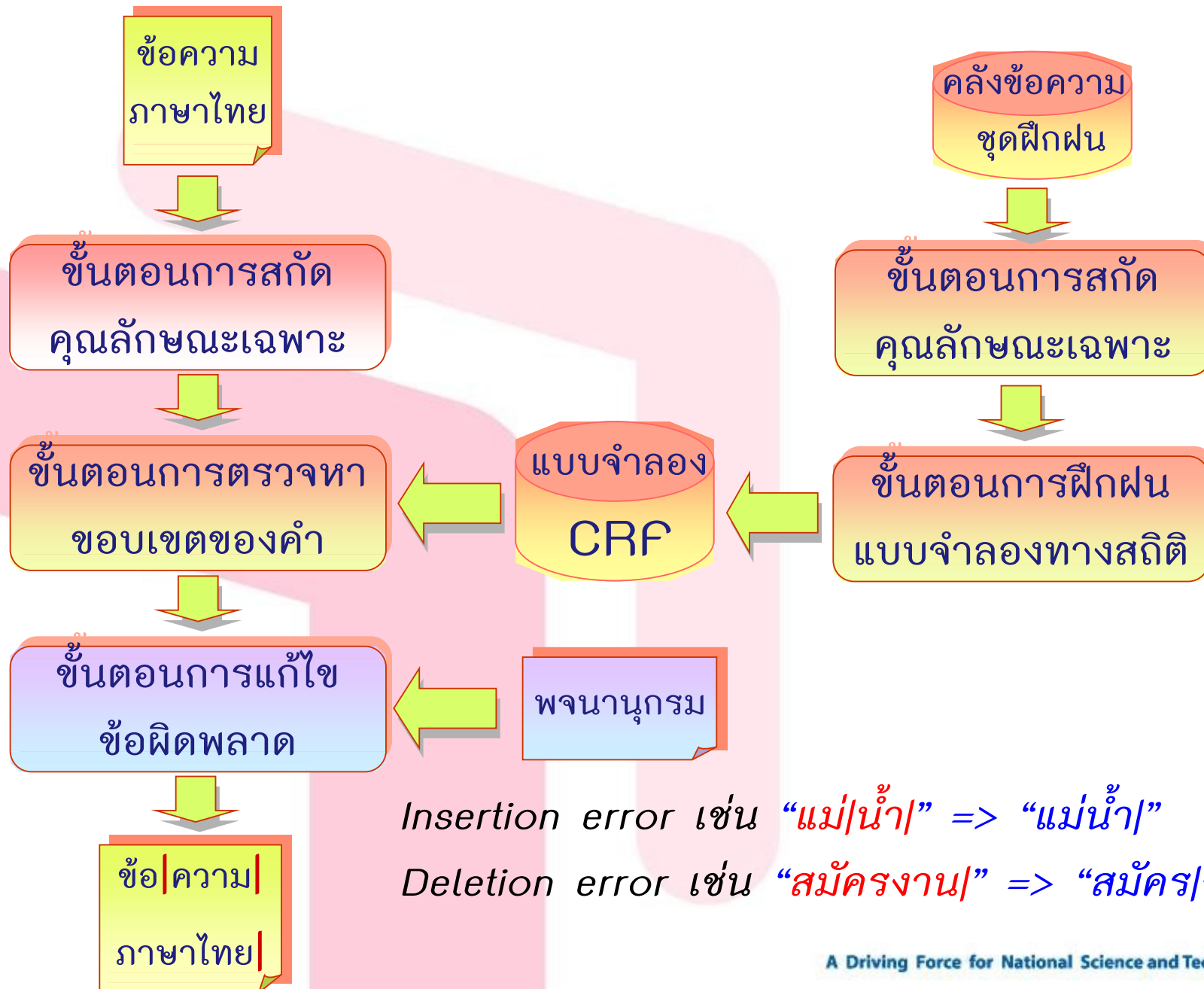
ขั้นตอนการสกัด
คุณลักษณะเฉพาะ

ประเภท	อักขระ
อักขระที่เป็นตัวสะกดได้	ก ข ช ค ง จ ฅ ช ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ ภ ม ย ร ล ว ศ ษ ส ฟ ฝ
อักขระที่เป็นตัวสะกดไม่ได้	ค จ ฅ ฝ ฬ ห ษ
สระที่เริ่มต้นคำไม่ได้	ะ ิ ี ึ ุ ึ ุ ำ ่า ำ ำ
สระที่เริ่มต้นคำได้	เ แ โ ใ ไ
วรรณยุกต์ไทย	่ ้ ๊ ๋ ๋
สัญลักษณ์ไทย	์ ์ ์ . ?
ตัวเลขไทย	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙
ตัวเลขอารบิก	0 1 2 3 4 5 6 7 8 9
สัญลักษณ์เงินบาท	฿
สัญลักษณ์พื้นฐานทั่วไป	? ^ " " @ # \$ % ? & * :
อักขระภาษาอังกฤษ	a-z A-Z
เครื่องหมายเว้นวรรค	วรรค
อักขระอื่นๆ	อักขรจีน ญี่ปุ่น เป็นต้น

B : เริ่มต้นคำ

I : ภายในคำ

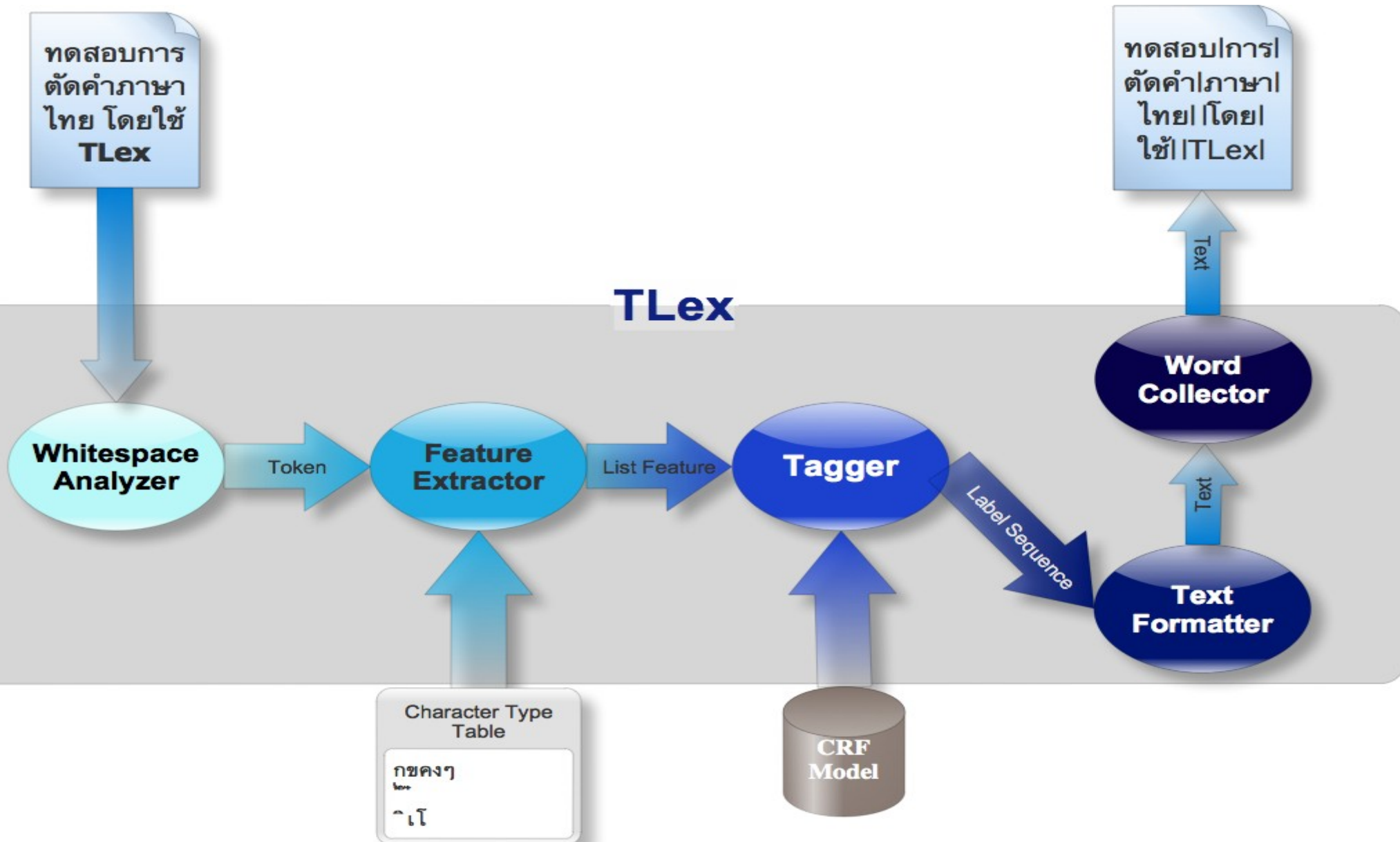
ระบบคววส์ (CUWS)



ระบบคววส์ (CUWS) ต่อ

- ระบบคววส์พัฒนาแบบจำลองด้วย *CRF++* ซึ่ง
เป็นซอฟต์แวร์ *opensource*

ระบบที่เลิกใช้ (Tlexs_10) ต่อ



ระบบทีเล็กซ์ (Tlexs_10) ต่อ

- ระบบทีเล็กซ์_10 พัฒนาแบบจำลองด้วย CRF++ ซึ่งเป็นซอฟต์แวร์ *opensource*

ระบบคำไทย (KAM Thai)

1. เทคนิคของ N-gram และ dictionary
2. โดยใช้โครงสร้างข้อมูล Trie ในการเก็บข้อมูล ซึ่งช่วยให้ทำงานได้เร็วมากขึ้น
3. โปรแกรมมีขนาดเล็ก
4. ข้อจำกัดเรื่องคำที่ไม่รู้จัก (Unknown word)

ตารางแสดงผลการแบ่งคำถูกต้องและผิดพลาด

ระบบ	คำที่แบ่งถูกต้อง		คำที่แบ่งผิดพลาด	
	known	unknown	known	unknown
ระบบ DHM	61,186	786	337	743
ระบบ CUWS	60,596	652	967	877
ระบบ KAM Thai	60,173	141	1,390	1,388
ระบบ TLexs_10	60,663	421	900	1,108

เปรียบเทียบคุณลักษณะเฉพาะ (ต่อ)

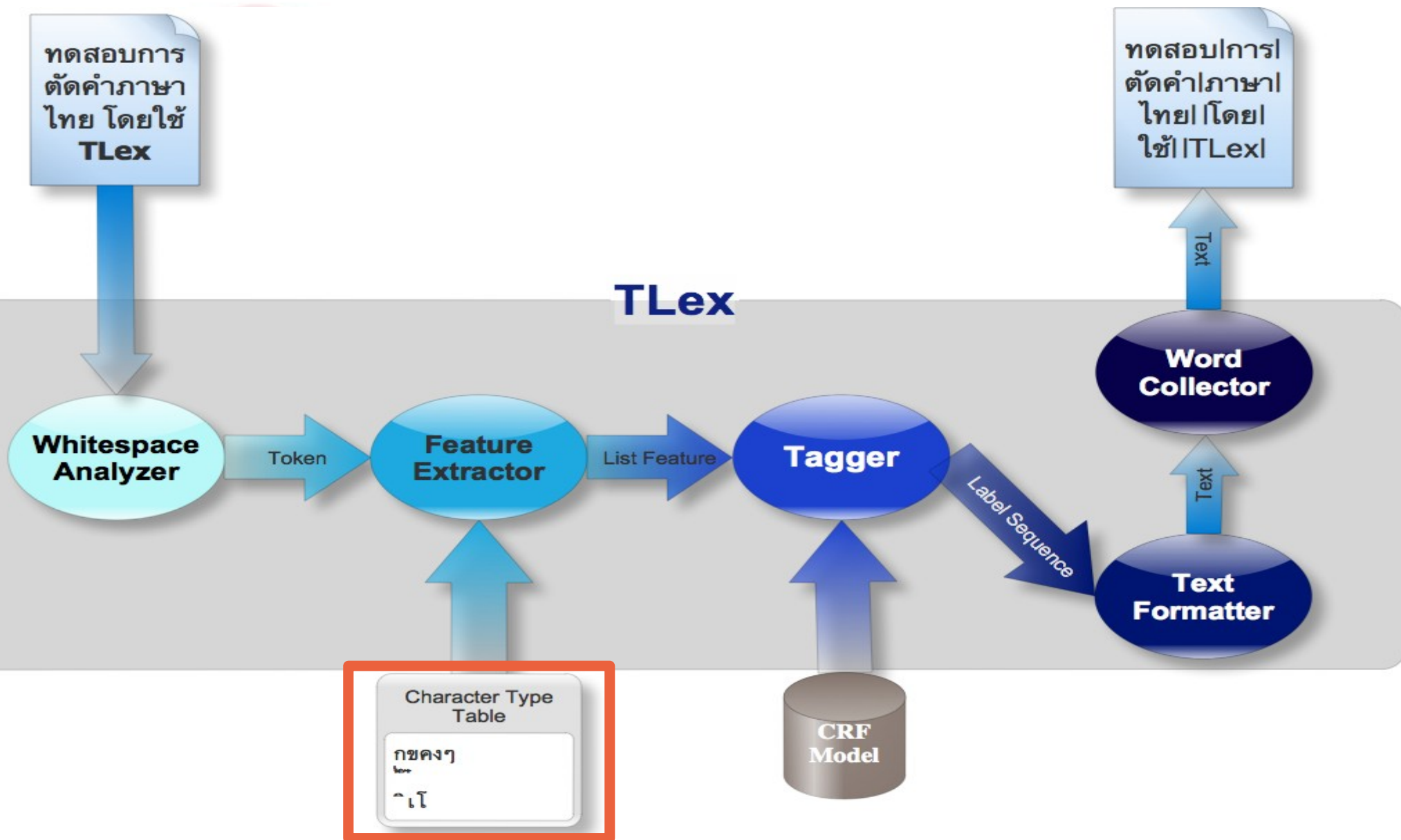
10

13

10

ประเภท	อักขระ	DHM	CUWS	Tlexs
ตัวเลข	0 1 2 3 4 5 6 7 8 9 ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙	-	-	1
ตัวเลขไทย	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙	1	1	-
ตัวเลขอารบิก	0 1 2 3 4 5 6 7 8 9	1	1	-
สัญลักษณ์เงินบาท	฿	-	1	-
เครื่องหมายอัฒประกาศ	'-' "-"	-	-	1
สัญลักษณ์พื้นฐานทั่วไป	? ^ " " @ # \$ % ? & * :	1	1	-
อักขระภาษาอังกฤษ	a-z A-Z	1	1	-
เครื่องหมายวรรค		1	1	1
อักขระอื่นๆ	อักษรจีน ญี่ปุ่น เป็นต้น	1	1	1 ₁₈

ระบบที่เลิกใช้ (TLexs_256)



คุณลักษณะเฉพาะของระบบที่เล็กซ์_256

		0	@	P	'	p				ฐ	ภ	๕๕	เ	๐
	!	1	A	Q	a	q			ก	ท	ม	เ	แ	๑
	"	2	B	R	b	r			ข	ฅ	ย	า	โ	๒
	#	3	C	S	c	s			ช	ณ	ร	ำ	ใ	๓
	\$	4	D	T	d	t			ค	ค	ฤ	เ	ไ	๔
	%	5	E	U	e	u			ก	ค	ถ	เ	า	๕
	&	6	F	V	f	v			ฅ	ถ	ภ	เ	า	๖
	'	7	G	W	g	w			ง	ท	ว	เ	เ	๗
	(	8	H	X	h	x			จ	ธ	ศ	เ	เ	๘
	)	9	I	Y	i	y			ฉ	น	ษ	เ	เ	๙
	*	:	J	Z	j	z			ช	บ	ส		เ	๐
	+	;	K	[	k	{			ช	ป	ห		เ	๑
	,	<	L	\	l				ณ	ผ	พ		เ	
	-	=	M	]	m	}			ญ	ฝ	อ		เ	
	.	>	N	^	n	~			ญ	พ	ฮ		เ	
	/	?	O	-	o				ญ	ฟ	า	฿	๐	

ตารางแสดงผลการแบ่งคำถูกต้องและผิดพลาด

ระบบ	คำที่แบ่งถูกต้อง		คำที่แบ่งผิดพลาด	
	known	unknown	known	unknown
ระบบ DHM	61,186	786	337	743
ระบบ CUWS	60,596	652	967	877
ระบบ KAM Thai	60,173	141	1,390	1,388
ระบบ TLexs_10	60,663		900	1,108
ระบบ TLexs_256	60,928		635	837

17.73%

ตารางเปรียบเทียบค่าต่างๆ ของทั้ง 5 ระบบ

ระบบ	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความเหวี่ยง
ระบบ DHM	98.22	97.67	97.95
ระบบ CUWS	98.10	97.58	97.84
ระบบ KAM Thai	97.97	96.42	97.20
ระบบ TLexs_10	95.52	95.92	95.72
ระบบ TLexs_256	97.34	97.22	97.28

สรุปผลการทดลอง

- คุณลักษณะเฉพาะของอักขระเป็นปัจจัยสำคัญในระบบแบ่งคำที่อาศัยการเรียนรู้ด้วยเครื่อง
- การเพิ่มจำนวนคุณลักษณะเฉพาะของอักขระสามารถช่วยให้ระบบที่เล็กลง แบ่งคำที่ไม่รู้จักได้ดีขึ้น
- คุณลักษณะเฉพาะไม่ใช่ปัจจัยเดียวที่ช่วยให้เครื่องมือแบ่งคำดีขึ้น

ข้อเสนอแนะ

- ทำการทดลองเพิ่มเพื่อหาจำนวนและประเภทของคุณลักษณะเฉพาะที่เหมาะสม
- หาปัจจัยอื่น ที่มีผลต่อการเพิ่มประสิทธิภาพของระบบแบ่งคำภาษาไทย

Q & A ขอบคุณค่ะ

<http://www.hlt.nectec.or.th/best/>