50307301 : สาขาวิชาวิทยาการคอมพิวเตอร์
คำสำคัญ : คำกำกวม, การตัดคำไทย, บริบทรอบข้าง, ความน่าจะเป็นแบบเบย์

กรรณิกา บุญเกษม : การเพิ่มประสิทธิภาพการตัดคำภาษาไทยโดยการพิจารณาบริบทรอบข้าง. อาจารย์ที่ปรึกษาวิทยานิพนธ์ : อ.ดร.ทัศนวรรณ ศูนย์กลาง. 151 หน้า.

การตัดคำในภาษาไทยปัญหาหลักที่พบคือ 1.ปัญหาความกำกวม 2.ปัญหาคำศัพท์ที่ไม่ปรากฏในพจนานุกรม วัตถุประสงค์ของการวิจัยครั้งนี้เพิ่มประสิทธิภาพการดัดคำภาษาไทยโดยมุ่งเน้นไปที่การแก้ปัญหาข้อความกำกวม ด้วยวิธีการเทียบคำยาวที่สุดร่วมกับการใช้กฎไวยากรณ์ทางภาษา โดยทำการทดลองเพื่อเปรียบเทียบจำนวนคำบริบทรอบข้างที่มีผลต่อประสิทธิภาพในการตัดคำ ซึ่งจำนวนคำบริบทรอบข้างจะพิจารณาทั้งคำด้านหน้าและหลังไม่เกิน 5 คำ ด้วยเทคนิคความน่าจะเป็นแบบเบย์ (Naive Bays) ในการทดลองใช้ชุดทดสอบจำนวน 2 ชุดทดสอบดังนี้ ชุดทดสอบจำนวน 500,000 คำ จากเอกสาร 5 ประเภท และชุดทดสอบแบบไม่แยกประเภทเอกสารจำนวน 100,000 คำ

จากผลการทดลองพบว่าจำนวนคำบริบทรอบข้าง 1 คำด้านหน้า 1 คำด้านหลัง (1-0-1) เป็นรูปแบบจำนวนคำบริบทรอบข้างที่เหมาะสมที่สุดที่สามารถให้ค่าความถูกต้องในการตัดคำมากที่สุดโดยไม่ขึ้นกับเอกสารประเภทใดประเภทหนึ่ง โดยให้ค่า F-measure ประมาณโดยเฉลี่ย 90% และให้ค่าความถูกต้องในการตัดคำเฉพาะการแก้ปัญหาความกำกวมประมาณ 87% โดยสามารถทำให้ค่าความถูกต้องเพิ่มขึ้นอย่างน้อยร้อยละ 3, 8 และ 10 เมื่อเปรียบเทียบกับกรณีไม่พิจารณาบริบทรอบข้าง โปรแกรม LexTo และโปรแกรม SWATH ตามลำดับ

จากผลการทดลองแสดงให้เห็นว่าการพิจารณาบริบทคำรอบข้าง 1 คำหน้า 1 คำหลัง เท่านั้นสามารถช่วยเพิ่มประสิทธิภาพการดัดคำภาษาไทยให้สูงขึ้นได้

---

ภาควิชาคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ปีการศึกษา 2554

ลายมือชื่อนักศึกษา.....

ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์ .....

50307301 : MAJOR : COMPUTER SCIENCE
KEY WORD : AMBIGUOUS WORD, THAI WORD SEGMENTATION, SURROUNDING CONTEXT, NAIVE BAYS
KANNIKA BOONKASEAM : IMPROVING THAI WORLD SEGMENTATION BY SURROUNDING CONTEXT. THESIS ADVISOR : Ph.D, TASANAWAN SOONKLANG.
151 pp.

There are two major problems in Thai word segmentation including ambiguous words and unknown words. This research aims to improve the performance of Thai word segmentation, focusing on the ambiguous words. The longest matching and rule-based algorithm approach are used for segmentation as a baseline model. The experiment presents a comparative study on the numbers of surrounding context within 5-word window sizes using Naive Bays algorithm.

The performance evaluation are experimented by using the two main corpuses, including approximately 500,000 words from five categories of documents, and 100,000 words from non-category document. The experiment results show that using one word in the back and one word in the front provides the best performance regardless of the type of document, with the F-measure approximately 90% and the accuracy in solving the ambiguity problem is approximately 87%. Comparing to the our baseline models and two well-known Thai word segmentation programs, called LexTo and SWATH, the performance (measured by F-measure) can be improved by increasing up to 3%, 8%, and 10% , respectively.

The experiment results show only one word in the back and one word in the front can be used to improve the accuracy of Thai word segmentation.

Department of Computing Graduate School, Silpakorn University Academic Year 2011
Student's signature ..... Kannika Boonkaseam .....
Thesis Advisor's signature .....

# กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลงได้ด้วยความช่วยเหลืออย่างดียิ่งของ อ.ดร.ทัศนวรรณ ศูนย์กลาง อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านให้คำปรึกษา แนะนำ ช่วยตรวจทาน แก้ไข งานวิจัยด้วยความเอาใจใส่เป็นอย่างดี และจุดประกายให้ผู้วิจัยทำงานงานวิจัยนี้

ผู้วิจัยขอขอบพระคุณ อ.ดร.สุนีย์ พงษ์พินิจภิญโญ ประธานกรรมการตรวจสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ อ.ดร.รัชฎา คงคะจันทร์ และ อ.ดร.คฑา ประดิษฐ์วงศ์ กรรมการผู้ทรงคุณวุฒิ ที่ช่วยให้งานวิจัยฉบับนี้สำเร็จสมบูรณ์ไปได้ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณมารดา ที่ให้ความรัก ความเข้าใจ ความเอาใจใส่ ให้โอกาสและสนับสนุนทุกด้าน รวมทั้งเป็นกำลังใจให้แก่ผู้วิจัยตลอดมา ทำให้งานวิจัยนี้สำเร็จด้วยดี