งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาระบบการรู้จำและการจำแนกประเภทของชื่อเฉพาะภาษาไทยโดยใช้แนวทางแบบลูกผสม (hybrid approach) โดยแนวทางดังกล่าวจะแบ่งออกเป็นสองส่วนคือส่วนที่เป็นระบบทางสถิติและส่วนที่เป็นระบบกฎ

สำหรับส่วนของระบบทางสถิตินั้นจะใช้วิธีทางสถิติร่วมกับโลคอลแมกซ์อัลกอริทึมเพื่อคัดเลือกกลุ่มพยางค์ที่อาจเป็นชื่อเฉพาะออกมา ซึ่งวิธีการทางสถิติที่ใช้ในการวัดความสัมพันธ์ระหว่างพยางค์ในที่นี้มี 5 วิธี ได้แก่ การใช้ค่ามิวชวลอินฟอร์เมชัน ค่าไคกำลังสอง ค่าคิวบิกแอสโซซิเอชันเรโช ค่าล็อกไลค์ลิฮูด และค่ามิวชวลเอ็กซ์เป็กเตชันนั้น ผลพบว่าวิธีที่ใช้ค่ามิวชวลเอ็กซ์เป็กเตชันร่วมกับการใช้โลคอลแมกซ์ อัลกอริทึม ในการรู้จำชื่อเฉพาะนั้นให้อัตราการรู้จำได้ผลดีที่สุด แต่วิธีดังกล่าวก็มีข้อเสียตรงที่ใช้เวลาในการประมวลผลที่นานเกินไป ทำให้ในงานวิจัยนี้จะใช้วิธีทางสถิติที่ให้ผลอัตราการรู้จำดีรองลงมา นั่นคือ การใช้ค่ามิวชวลอินฟอร์เมชันร่วมกับการใช้โลคอลแมกซ์ อัลกอริทึม จากนั้นเมื่อได้ชื่อเฉพาะที่เลือกมาด้วยวิธีการทางสถิติแล้ว จะเข้าสู่ส่วนที่เป็นระบบกฎ ซึ่งระบบถูกเขียนขึ้นโดยอิงกับหลักฐานที่ได้จากบริบทภายใน เช่น คำนำหน้าชื่อ และใช้บริบทข้างเคียง เช่น คำปรากฏร่วม เพื่อช่วยในการรู้จำและจำแนกประเภทของชื่อเฉพาะ และจากการทดสอบพบว่าระบบกฎที่สร้างขึ้นสามารถจำแนกประเภทของชื่อเฉพาะโดยให้อัตราการรู้จำ (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

This study aims to develop a Thai named entity recognition and classification system using a hybrid approach. The system is composed of two parts, which are statistical part and rule part.

Statistical part is used for extracting named entity candidates. Localmaxs algorithm and the statistical method are used for measuring associations between syllables. Five statistical methods namely Mutual Expectation, Mutual Information, Chi-square, Cubic Association ratio and Loglikelihood are tested in this part. Mutual Expectation combined with Localmaxs algorithm yields the best result, but this method uses much more times than other methods. Therefore, Mutual Information, which is the second best statistical method, combined with Localmaxs algorithm is used for extracting a chunk of syllables as a candidate of named entity. On the second part, named entity candidates will be recognized and classified by linguistic rules which are manually crafted. Internal evidence, i.e. title names, and external evidence, i.e. collocate words, are used in these rules. The system can recognize and classify named entities with the recognition rate (F-measure), precision and recall rates at 69.15% , 54.00% and 96.12% for person names, 62.95% , 47.60% and 92.93% for organization names, 38.87% , 31.67% and 50.32% for location names.