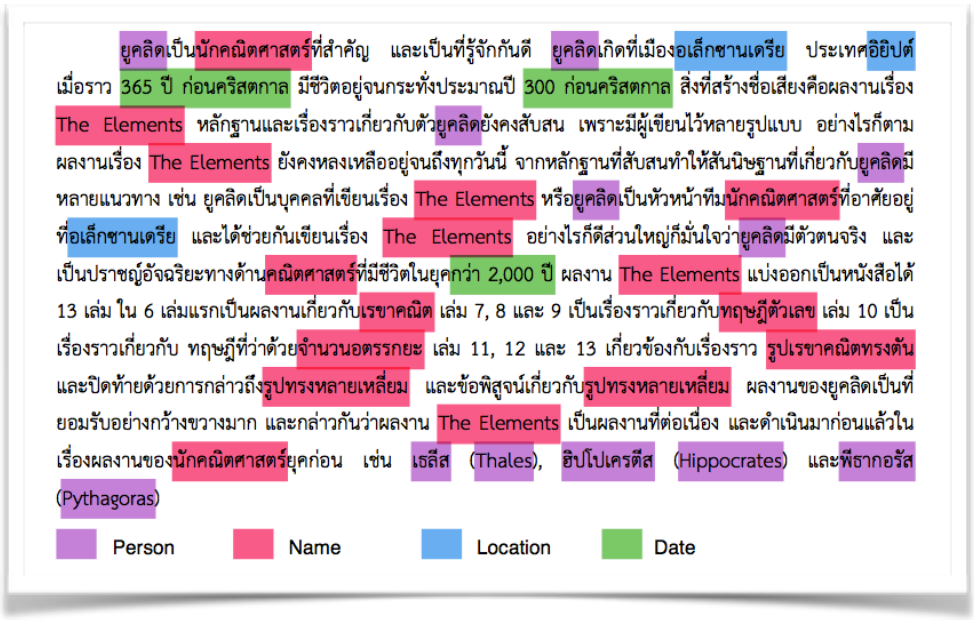
การรู้จำชื่อเฉพาะภาษาไทย

Thai Named Entity Recognition (NER)



การรู้จำชื่อเฉพาะ (NER)

ปัจจุบันอินเทอร์เน็ตเป็นสิ่งสำคัญในชีวิตประจำวันโดยกิจกรรมหนึ่งที่คนนิยมใช้จากอินเทอร์เน็ตคือการสืบค้นข้อมูล แต่การสืบค้นชื่อเฉพาะอาจได้ผลลัพธ์ไม่เป็นตามที่ต้องการเนื่องจากเครื่องมือไม่สามารถแยกคำเฉพาะจากคำนามทั่วไปได้ เช่น เมื่อสืบค้นคำว่า “หนังเพื่อนสนิท” ซึ่งเป็นชื่อเฉพาะที่เป็นชื่อภาพยนตร์ อาจได้ผลลัพธ์เป็น“หนังเพื่อนสนิทชวนดู”ได้ ซึ่งไม่ตรงตามที่ต้องการ นอกจากนี้ปัญหาที่เกี่ยวกับคำเฉพาะยังส่งผลถึงการแปลคำอีกด้วย เพราะเมื่อชื่อเฉพาะถูกแปลอาจทำให้ความหมายผิดเพี้ยนได้ เช่น “Sizzler” ซึ่งเป็นชื่อร้านอาหาร ถ้าเครื่องมือไม่รู้ว่า “Sizzler” เป็นชื่อเฉพาะ “Sizzler” จะถูกแปลว่า “สิ่งที่ร้อนจัด” แทน ดังนั้นการรู้จำชื่อเฉพาะเป็นการรู้จำชื่อที่เฉพาะเจาะจงโดยสกัดชื่อเฉพาะออกจากคำนามทั่วไปทำให้การวิเคราะห์ข้อมูลทางภาษาได้ถูกต้องแม่นยำขึ้น

ประโยชน์การรู้จำชื่อเฉพาะ

1 การสืบค้น

ผลลัพธ์ในการสืบค้นเป็นตามที่ต้องการมากขึ้น

2 การแปล

คำเฉพาะจะไม่ถูกแปลจึงทำให้ลดความผิดเพี้ยนจากการแปล

3 การระบุโครงสร้างประโยค

การรวมคำที่เป็นคำเฉพาะทำให้การระบุโครงสร้างของประโยคถูกต้องขึ้น

ลักษณะชื่อเฉพาะในภาษาอังกฤษและภาษาไทย

คำเฉพาะในภาษาอังกฤษ

ในภาษาอังกฤษ มีการเว้นวรรคคำทำให้รู้ขอบเขตของคำ รวมถึงมีตัวพิมพ์ใหญ่ในการบ่งบอกถึงชื่อเฉพาะต่างๆ

คำเฉพาะในภาษาไทย

ในภาษาไทย ไม่มีลักษณะที่บ่งบอกขอบเขตของคำ รวมถึงลักษณะที่บ่งบองถึงชื่อเฉพาะ ทำให้การรู้จำชื่อเฉพาะภาษาไทยยุ่งยากกว่าภาษาอังกฤษ

ชนิดของคำเฉพาะ

โดยส่วนมากจะแบ่งเป็น 6 ชนิด ได้แก่

วัน/เดือน/ปี เช่น วันจันทร์ เมื่อวาน 30 เมษายน 2560 ฯลฯ

สถานที่ เช่น ประเทศไทย พารากอน ฯลฯ

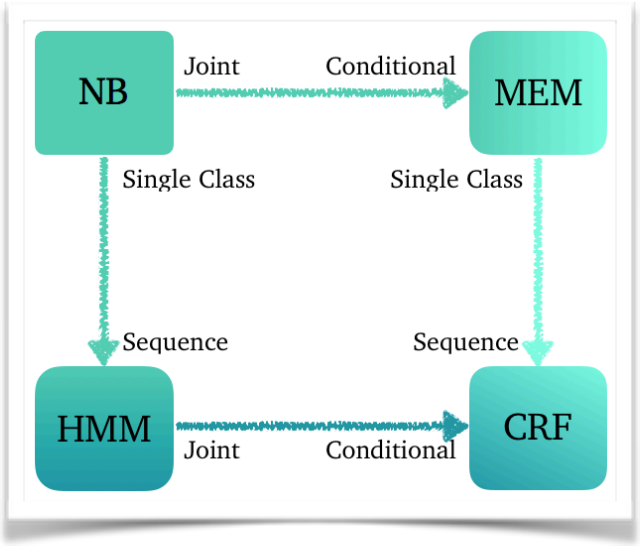
ชื่อองค์กร เช่น องค์การบริหารส่วนตำบล อบต. ฯลฯ

ชื่อคน เช่น พล.อ.ประยุทธ์ จันทร์โอชา ฯลฯ

เวลา เช่น 18:00 น. 30 นาที ฯลฯ

ชื่อเฉพาะอื่นๆ เช่น ชื่อตำแหน่งงาน ชื่อสินค้า ฯลฯ

แบบจำลองความน่าจะเป็นในการรู้จำชื่อเฉพาะ



แบบจำลองความน่าจะเป็นถูกนำมาใช้ในการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) ทั้งในการตัดคำ (Word Segmentation) การกำกับหมวดคำ (Part of Speech Tagging) โดยเฉพาะการรู้จำชื่อเฉพาะ (Named Entity Recognition : NER) เพื่อแยกชื่อเฉพาะออกจากคำนามทั่วไป

แบบจำลองความน่าจะเป็นมีหลายรูปแบบ เช่น Naïve Bayes (NB) Hidden Markov Model (HMM) Maximum Entropy Model (MEM) และ Conditional Random Field (CRF) ฯลฯ

NB เป็นแบบจำลองเริ่มแรกที่นำมาใช้กับ NLP โดยนำทฤษฎีบทของเบยส์มาใช้ NB เป็นแบบจำลองแบบ Generative model ซึ่งเป็นแบบจำลองที่ใช้ความน่าจะเป็นร่วม (Joint Probability) แต่ยังมีข้อบกพร่องตรงที่ NB มีการสมมติลักษณะข้อมูลนำเข้าไว้ ถ้าข้อมูลมีลักษณะไม่ตรงตามที่สมมติไว้ก็ไม่ควรใช้ NB และ NB ถูกใช้ในการทำนายกลุ่มที่เป็นกลุ่มเดียว ต่อมามีการพัฒนา NB เป็น HMM ซึ่งเป็นแบบจำลองที่ใช้ทำนายลำดับของกลุ่ม ต่อมามีการพัฒนา NB เป็น MEM ซึ่งไม่มีข้อจำกัดของข้อมูลนำเข้าและแบบจำลองเป็นแบบ Discriminative Model ซึ่งเป็นแบบจำลองที่ใช้ความน่าจะเป็นแบบมีเงื่อนไข (Coonditional Probability) ส่วน CRF คือแบบจำลองความน่าจะเป็นที่พัฒนามาจาก NB HMM และ MEM โดยแบบจำลองที่ได้จะเป็นแบบ Discriminative Model เพื่อทำนายผลที่ลักษณะเป็นลำดับ โดย CRF สามารถใช้ข้อมูลนำเข้าที่ไม่มีข้อจำกัดของข้อมูลและทำนายผลที่มีลักษณะเป็นลำดับได้

Reference

1. Klinger R., Tomanek K.. , *Classical probabilistic models and conditional random fields.* , 2007 Technical Report TR07-2-013. Department of Computer Science, Dortmund University of Technology ISSN 1864-4503

2. N. Tirasaroj and W. Aroonmanakun, "Thai named entity recognition based on conditional random fields," *2009 Eighth International Symposium on Natural Language Processing*, Bangkok, 2009, pp. 216-220.