

# การรู้จำชื่อเฉพาะภาษาไทย

## Thai Named Entity Recognition (NER)

ยูคลิดเป็นนักคณิตศาสตร์ที่สำคัญ และเป็นที่รู้จักกันดี ยูคลิดเกิดที่เมืองเอเล็กซานเดรีย ประเทศอียิปต์ เมื่อราว 365 ปีก่อนคริสตกาล มีชีวิตอยู่จนกระทั่งประมาณปี 300 ก่อนคริสตกาล สิ่งที่เขาสร้างชื่อเสียงคือผลงานเรื่อง The Elements หลักฐานและเรื่องราวเกี่ยวกับตัวยูคลิดยังคงสับสน เพราะมีผู้เขียนไว้หลายรูปแบบ อย่างไรก็ตาม ผลงานเรื่อง The Elements ยังคงหลงเหลืออยู่จนถึงทุกวันนี้ จากหลักฐานที่สับสนทำให้สันนิษฐานเกี่ยวกับยูคลิดมีหลายแนวทาง เช่น ยูคลิดเป็นบุคคลที่เขียนเรื่อง The Elements หรือยูคลิดเป็นหัวหน้าทีมนักคณิตศาสตร์ที่อาศัยอยู่ที่เอเล็กซานเดรีย และได้ช่วยกันเขียนเรื่อง The Elements อย่างไรก็ตามก็มีส่วนที่มั่นใจว่ายูคลิดมีตัวตนจริง และเป็นปราชญ์อัจฉริยะทางด้านคณิตศาสตร์ที่มีชีวิตในยุคกว่า 2,000 ปี ผลงาน The Elements แบ่งออกเป็นหนังสือได้ 13 เล่ม ใน 6 เล่มแรกเป็นผลงานเกี่ยวกับเรขาคณิต เล่ม 7, 8 และ 9 เป็นเรื่องราวเกี่ยวกับทฤษฎีตัวเลข เล่ม 10 เป็นเรื่องราวเกี่ยวกับ ทฤษฎีที่ว่าด้วยจำนวนอตรรกยะ เล่ม 11, 12 และ 13 เกี่ยวข้องกับเรื่องราว รูปเรขาคณิตทรงตัน และปิดท้ายด้วยการกล่าวถึงรูปทรงหลายเหลี่ยม และข้อพิสูจน์เกี่ยวกับรูปทรงหลายเหลี่ยม ผลงานของยูคลิดเป็นที่ยอมรับอย่างกว้างขวางมาก และกล่าวกันว่าผลงาน The Elements เป็นผลงานที่ต่อเนื่อง และดำเนินมาก่อนแล้วในเรื่องผลงานของนักคณิตศาสตร์ยุคก่อน เช่น เธลีส (Thales), ฮิปโปเครติส (Hippocrates) และพีทาโกรัส (Pythagoras)

Person Name Location Date

### ลักษณะชื่อเฉพาะในภาษาอังกฤษและภาษาไทย

วงกลมใหญ่ก็มั่นใจว่ายูคลิดเขียน The Elements และเรื่องราวเกี่ยวกับทฤษฎีต่างๆ

#### คำเฉพาะในภาษาอังกฤษ

ในภาษาอังกฤษ มีการเว้นวรรคคำทำให้รู้ขอบเขตของคำ รวมถึงมีตัวพิมพ์ใหญ่ในการบ่งบอกถึงชื่อเฉพาะต่างๆ

วงกลมใหญ่ก็มั่นใจว่ายูคลิดเขียน The Elements และเรื่องราวเกี่ยวกับทฤษฎีต่างๆ

#### คำเฉพาะในภาษาไทย

ในภาษาไทย ไม่มีลักษณะที่บ่งบอกขอบเขตของคำ รวมถึงลักษณะที่บ่งบอกถึงชื่อเฉพาะ ทำให้การรู้จำชื่อเฉพาะภาษาไทยยุ่งยากกว่าภาษาอังกฤษ

## การรู้จำชื่อเฉพาะ (NER)

ปัจจุบันอินเทอร์เน็ตเป็นสิ่งสำคัญในชีวิตประจำวันโดยกิจกรรมหนึ่งที่คุณนิยมใช้จากอินเทอร์เน็ตคือการสืบค้นข้อมูล แต่การสืบค้นชื่อเฉพาะอาจได้ผลลัพธ์ไม่เป็นตามที่ต้องการเนื่องจากเครื่องมือไม่สามารถแยกคำเฉพาะจากคำนามทั่วไปได้ เช่น เมื่อสืบค้นคำว่า “หนังสือเพื่อนสนิท” ซึ่งเป็นชื่อเฉพาะที่เป็นชื่อภาพยนตร์ อาจได้ผลลัพธ์เป็น “หนังสือเพื่อนสนิทชวนดู” ได้ ซึ่งไม่ตรงตามที่ต้องการ นอกจากนี้ปัญหาเกี่ยวกับคำเฉพาะยังส่งผลถึงการแปลคำอีกด้วย เพราะเมื่อชื่อเฉพาะถูกแปลอาจทำให้ความหมายผิดเพี้ยนได้ เช่น “Sizzler” ซึ่งเป็นชื่อร้านอาหาร ถ้าเครื่องมือไม่รู้ว่า “Sizzler” เป็นชื่อเฉพาะ “Sizzler” จะถูกแปลว่า “สิ่งที่ร้อนจัด” แทน ดังนั้นการรู้จำชื่อเฉพาะเป็นการรู้จำชื่อที่เฉพาะเจาะจงโดยสกัดชื่อเฉพาะออกจากคำนามทั่วไปทำให้การวิเคราะห์ข้อมูลทางภาษาได้ถูกต้องแม่นยำขึ้น

### ประโยชน์การรู้จำชื่อเฉพาะ

1

#### การสืบค้น

ผลลัพธ์ในการสืบค้นเป็นตามที่ต้องการมากขึ้น

2

#### การแปล

คำเฉพาะจะไม่ถูกแปลจึงช่วยลดความผิดเพี้ยนจากการแปล

3

#### การระบุโครงสร้างประโยค

การรวมคำที่เป็นคำเฉพาะทำให้การระบุโครงสร้างของประโยคถูกต้องขึ้น

# ชนิดของคำเฉพาะ

โดยส่วนมากจะแบ่งเป็น 6 ชนิด ได้แก่

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

สถานที่ เช่น ประเทศไทย พาราгон ฯลฯ

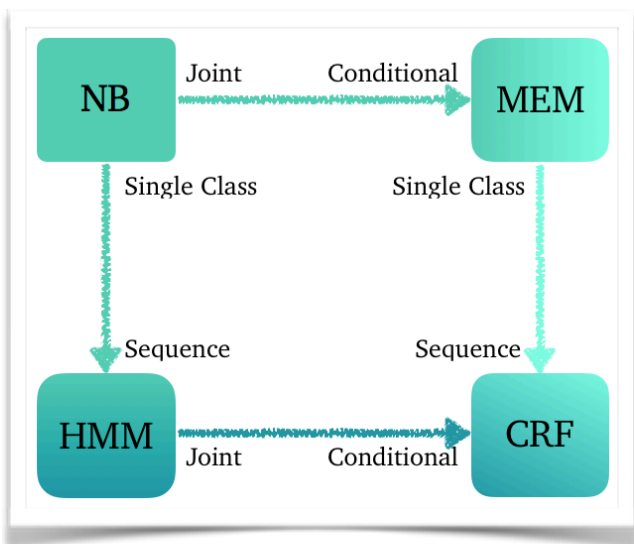
ชื่อองค์กร เช่น องค์กรบริหารส่วนตำบล อบต. ฯลฯ

ชื่อคน เช่น พล.อ.ประยุทธ์ จันทร์โอชา ฯลฯ

เวลา เช่น 18:00 น. 30 นาที ฯลฯ

ชื่อเฉพาะอื่นๆ เช่น ชื่อตำแหน่งงาน ชื่อสินค้า ฯลฯ

## แบบจำลองความน่าจะเป็นในการรู้จำชื่อเฉพาะ



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

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

NB เป็นแบบจำลองเริ่มแรกที่นำมาใช้กับ NLP โดยนำทฤษฎีบทของเบย์ส์มาใช้ 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 (Conditional 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.