KM Annotation Guideline

Nathanon Theptakob Nov2022

เป้าหมาย

การทำเครื่องหมาย (annotate) บนคำที่ถือว่าเป็น entity ทุกคำ และทำการเชื่อมโยง entities ที่เหมือนกันในเชิงความหมาย (coreference) เข้าด้วยกัน ทั้งภายในเอกสารเดียวกัน (Within-document coreference annotation) และต่างเอกสาร (Cross-document coreference annotation)

ขั้นตอนการทำเครื่องหมาย

1. ลากเมาส์คลุมดำครอบคำที่ต้องการเพื่อทำเครื่องหมาย ส่วนที่ถูกครอบพร้อมกันนั้นจะถูกเรียกว่าเป็น entity เดียวกัน โดยให้ทำเครื่องหมายเฉพาะส่วนที่เป็น entity เท่านั้น ไม่ต้องทำส่วนที่เป็น event
   ยกตัวอย่างเช่น
   สมชายเป็นนักแข่งรถมือหนึ่งที่เข้าร่วมแข่งขันในงานแรลลี่ งานนี้มีผู้เข้าแข่งขันถึง 100 คน
   ให้ทำเครื่องหมายเฉพาะส่วนที่เป็นสีน้ำเงิน (entity) เท่านั้น ไม่ต้องทำเครื่องหมายในส่วนสีแดง (event)
2. หลังจากทำเครื่องหมาย entity ทั้งหมดแล้ว ให้ลากเมาส์จาก entity ที่ถูกทำเครื่องหมายแล้วไปหา entity ที่ถูกทำเครื่องหมายอีกตัวหนึ่งเพื่อเป็นการเชื่อมโยงว่า 2 entities นี้เหมือนกันในเชิงความหมาย ในขั้นตอนนี้ให้นับว่าคำสรรพนาม (pronoun) ที่ใช้แทนความหมายของ entity ว่าเหมือนกันในเชิงความหมายกับชื่ออื่น ๆ ของ entity นั้น ๆ ด้วย
   ยกตัวอย่างเช่น
   สมชายเป็นนักแข่งรถมือหนึ่งที่เข้าร่วมแข่งขันในงานแรลลี่ เขาต้องสู้กับผู้เข้าแข่งขันถึง 100 คน
   ทั้ง entity “สมชาย” “นักแข่งรถมือหนึ่ง” และ “ผู้เข้าแข่งขัน” เหมือนกันในเชิงความหมาย โดยให้รวมคำว่า “เขา” เข้าไปในกลุ่มนี้ด้วย เพราะคำว่า “เขา” ในที่นี้สื่อถึงสมชาย
   สามารถเข้าไปดูตัวอย่างการทำเครื่องหมายและการเชื่อมโยงได้ที่ INCEpTION - Projects (tu-darmstadt.de)
   Username, Password: Demo
   Project name: Thai_test
3. เมื่อทำเครื่องหมายและเชื่อมโยง entity ที่เหมือนกันในเชิงความหมายเรียบร้อยแล้ว ให้ export ผลการทำเครื่องหมายออกมาใน CoNLL 2012 format และอัพโหลดขึ้นไปบน google drive ที่กำหนดให้

https://drive.google.com/drive/folders/1l49LSEozzXg-YbrR7YaMx908xeQiw493?usp=sharing

4. สำหรับ entity ที่เหมือนกันในเชิงความหมายที่ปรากฏในต่างเอกสารกันให้จดแยกไว้ใน google sheet ที่กำหนดให้ https://docs.google.com/spreadsheets/d/13s2Da9YxZKcXW-XyHi7fJpK-TlBr_ytuYhLlzhSjtn4/edit?usp=sharing โดยมีลักษณะการเก็บข้อมูล (data format) ดังนี้
    a. เมื่อมี entity ชนิดใหม่ที่ไม่เคยเชื่อม cross-document coreference มาก่อน ให้เขียนชื่อของ entity นั้นไว้ในคอลัมน์แรก “Entity” โดยไม่จำเป็นต้องเป็นชื่อเต็มของ entity นั้น ๆ ก็ได้ (ใช้เป็นชื่อเล่นก็ได้)
    b. ในคอลัมน์ต่อ ๆ ไป จะเป็นชื่อ document ที่เรานำมา annotate โดยข้อมูลในคอลัมน์จะมีลักษณะดังนี้

        <รหัส entity>D<หมายเลขเอกสาร>

        โดยรหัส entity และหมายเลขเอกสาร สามารถดูได้จากไฟล์ .conll ที่ export ออกมาในขั้นตอนที่ 3 ยกตัวอย่างเช่น

        “คณบดีคณะแพทยศาสตร์ศิริราชพยาบาล มหาวิทยาลัยมหิดล” จะใช้รหัส 1D1

        เมื่อใส่ใน google sheet จะลงข้อมูลว่า

        1D1 ในคอลัมน์ที่ตรงกับ txt_og_1.txt และแถวที่ตรงกับ “คณบดีศิริราช”

Note: ใน google sheet จะมีชีท “Example” สำหรับเป็นตัวอย่างในการ annotate อยู่แล้ว เมื่อ annotate จริงให้เข้าไปทำในชีท “Annotation”