ว.วิทย. มข. 41(1) 121-133 (2556) KKU Sci. J. 41(1) 121-133 (2013)

แนวคิดและความท้าทายในการพัฒนาการค้นคืนข้อมูลข้ามภาษาไทย-อังกฤษ

Cross Language (Thai-English) Information Retrieval: Concepts and Challenges

ไกรศักดิ์ เกษร[1]

บทคัดย่อ

ปัจจุบันเอกสารที่กระจายอยู่ทั่วไปในอินเทอร์เน็ตมีความหลากหลายทางด้านภาษา ข้อดีคือผู้ใช้มีข้อมูลที่หลากหลายและสามารถนำไปใช้ให้เป็นประโยชน์ต่อการศึกษาหรือการทำงาน นอกจากนี้ผู้ใช้ยังสามารถใช้ข้อมูลเหล่านี้มาตรวจสอบและยืนยันความถูกต้องซึ่งกันและกันได้ แต่ข้อเสียที่ตามมาคือผู้ใช้ต้องทำการค้นหาข้อมูลโดยใช้คำสำคัญที่เป็นภาษาเดียวกับเอกสารเท่านั้นจึงจะได้ผลลัพธ์ตามที่ต้องการ นอกจากนี้ผู้ใช้อาจจะไม่รู้จะใช้คำศัพท์ใดที่จะอธิบายถึงข้อมูลที่ตนเองต้องการได้ ซึ่งมีผลทำให้ระบบการค้นคืนข้อมูลไม่สามารถค้นหาเอกสารที่ผู้ใช้ต้องการได้อย่างถูกต้องแม่นยำ เนื่องจากลักษณะของคำศัพท์ในแต่ละภาษามีความหมายที่ไม่ตายตัว โดยปกติคำหนึ่งคำสามารถมีหลายความหมาย เรียกว่า “Polysemy” หรือคำหลายคำสามารถหมายถึงสิ่งเดียวกันเรียกว่า “Synonym” ระบบค้นหาสารสนเทศในปัจจุบันยังมีประสิทธิภาพต่ำในการแก้ปัญหาเหล่านี้ ด้วยความสำคัญของปัญหาดังกล่าวนักวิจัยจึงมีแนวคิดที่จะพัฒนาวิธีการค้นคืนสารสนเทศข้ามภาษา (Cross-Language Information Retrieval-CLIR) ขึ้น เพื่อช่วยให้ผู้ใช้ที่เป็นคนไทยได้ข้อมูลที่ตรงกับสิ่งที่ผู้ใช้ต้องการมากที่สุด ถึงแม้จะมีข้อจำกัดในเรื่องของภาษาหรือคำศัพท์ภาษาอังกฤษที่จะใช้ในคิวรี่ (Query) ก็ตาม แนวความคิดนี้ถือเป็นแนวโน้มใหม่สำหรับระบบค้นหาสารสนเทศ (Information Retrieval-IR) และมีศักยภาพสูงในการพัฒนาต่อยอดในเชิงพาณิชย์ให้กับระบบค้นหาข้อมูลเช่น Google หรือ Bing ได้ ในบทความนี้นำเสนอถึงแนวคิดของ CLIR และสรุปความท้าทายสำหรับนักวิจัยที่ต้องการจะสร้างระบบ CLIR สำหรับภาษาไทยและอังกฤษขึ้นมา

[1]ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร อ.เมือง จ.พิษณุโลก 65000
E-Mail: kraisakk@nu.ac.th

ABSTRACT

Documents on the Internet have been written using several languages. The benefit of this is those documents are useful for users to verify the information from different sources. However, users are not able to use a single language in a query to retrieve all relevant documents written in different languages. Moreover, some users do not know exactly what the keywords to be used in a query to retrieve desired documents. As a result, search engine cannot find the relevant documents effectively. In addition, a keyword can refer to many different concepts in the real world, so-called “Polysemy” or many keywords refer to one thing, so-called “Synonym”. They are two significant problems that decrease a search engine performance. Consequently, many researchers try to overcome the problem by developing the Cross-Language Information Retrieval (CLIR) system in order to retrieve documents written by different languages from using a single query. This idea is now a new trend of search engine and can be developed as a commercial product for a popular search engine e.g. Google or Bing. The article presents the concepts and ideas of CLIR and summary of the main challenges in this research area.

คำสำคัญ: ข้ามภาษา สองภาษา ระบบคืนคืนสารสนเทศ เครื่องแปลความหมาย

Keywords: Cross-language, Translingual, Bilingual, Information retrieval, Machine translation

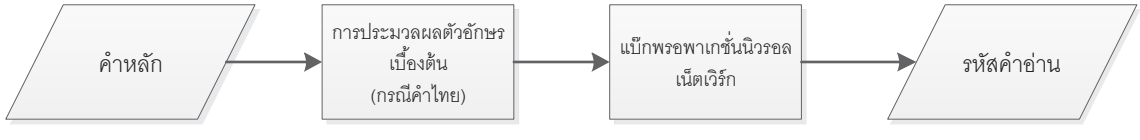
1. บทนำ

ข้อมูลบนอินเทอร์เน็ตมีความหลากหลายในด้านภาษาและประเภทของข้อมูล เช่น เป็นตัวอักษร รูปภาพ หรือวีดีโอ ด้วยเหตุนี้ทำให้การค้นหาข้อมูลให้ตรงความต้องการได้ยากขึ้น จึงทำให้ผู้ใช้ต้องทำการค้นหาข้อมูลโดยใช้คำสำคัญ (keyword) เป็นภาษาเดียวกับเอกสารเท่านั้น นอกจากนี้ปัญหายังเกิดกับผู้ใช้ที่ไม่ชำนาญในการใช้คำศัพท์ภาษาอื่นๆ ที่ไม่ใช่ภาษาประจำชาติของตนเอง เนื่องจากไม่รู้ว่าจะใช้คำศัพท์คำไหนที่ใช้อธิบายข้อมูลที่ต้องการค้นหาได้ถูกต้อง ทำให้ระบบค้นหาข้อมูล (search engine) ไม่สามารถค้นคืนข้อมูลได้ถูกต้องตามความต้องการของผู้ใช้ ดังนั้นจึงมีนักวิจัยหลายท่านได้พยายามพัฒนาเทคนิคที่จะรองรับกระบวนการที่จะค้นหาเอกสารที่ถูกเขียนขึ้นภาษาหนึ่งและสามารถถูกค้นหาโดยใช้คิวรี่อีกภาษาหนึ่งได้ ดังนั้นสรุปความหมายหรือนิยามของระบบค้นคืนข้อมูลข้ามภาษา (CLIR) คือ ระบบที่สามารถประมวลผลคิวรี่เพื่อค้นคืนข้อมูลเอกสารที่ถูกเขียนโดยใช้ภาษาที่ต่างจากคิวรี่ เพื่อให้ได้เอกสาร รูปภาพ หรือข้อมูลเสียงที่สอดคล้องกับคิวรี่ การค้นคืนสารสนเทศข้ามภาษานี้มีชื่อเรียกที่ใช้เรียกอยู่หลายคำ เช่น การค้นคืนสารสนเทศหลายภาษา (mutilingual information retrieval-MLIR (Hull and Grefenstette, 1996) หรือ การค้นคืนสารสนเทศแปลภาษา (translingual information retrieval) (Yang et al., 1998) เทคนิคในการพัฒนาระบบค้นคืนข้อมูลข้ามภาษานั้นสามารถ

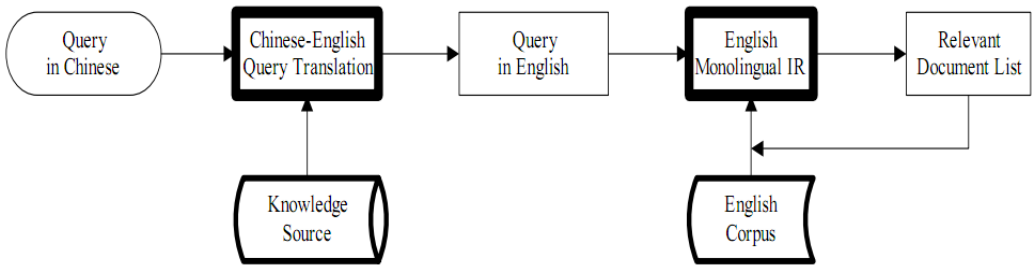
ทำได้หลายวิธี เทคนิคซึ่งประเทศต่าง ๆ ทั่วโลกได้มีการทำวิจัยและนำเสนอเทคนิคต่าง ๆ ในการค้นคืนข้อมูลข้ามภาษา ดังนั้นในบทความนี้จะขอยกตัวอย่างวรรณกรรมที่เกี่ยวข้องกับระบบ CLIR ในภาษาต่าง ๆ เช่น จีน ญี่ปุ่น และไทย เป็นต้น

2. การพัฒนาระบบ CLIR สำหรับภาษาต่าง ๆ

แนวคิดเกี่ยวกับการค้นคืนข้อมูลข้ามภาษานี้ถูกพัฒนาขึ้นเมื่อประมาณ 50 ปีที่แล้ว โดยเริ่มจาก Salton (1969) ได้พัฒนาระบบเพื่อค้นเอกสารข้ามภาษาระหว่างภาษาเยอรมันและภาษาอังกฤษและต่อมาได้พัฒนาระบบค้นเอกสารระหว่างภาษาอังกฤษและฝรั่งเศสในปี 1973 (Salton, 1973) นักวิจัยชื่อ Douglas (1997) ได้พยายามที่จะพัฒนาระบบค้นคืนเอกสารข้ามภาษาระหว่างภาษาสเปนและภาษาอังกฤษ Fredric (2001) พัฒนาระบบค้นเอกสารระหว่างภาษาอาราบิคและภาษาอังกฤษ Chen (2002) ทำการสร้างระบบค้นหาเอกสารระหว่างภาษาจีนและภาษาอังกฤษ Ampazis (2004) ได้นำเสนอการทำดัชนีข้ามภาษาโดยใช้วิธีการของ LSI (latent semantic indexing) (Deerwester et al., 1990) เพื่อทำการค้นหาข้อมูลจากปริภูมิเว็กเตอร์ (vector space model) และนำวิธีการจัดกลุ่มข้อมูลที่เรียกว่า SOM (self-organize maps) มาช่วยในการตีความของคิวรี่จากผู้ใช้ แต่วิธีการนี้อาจจะมีปัญหากับภาษาไทยเนื่องจากภาษาไทยเป็นภาษาที่ไม่มีช่องว่างระหว่างคำ เนื่องจาก LSI ทำดัชนีเป็นคำ ๆ ดังนั้นการประยุกต์ใช้กับภาษาไทยอาจจะต้องพัฒนาส่วนของการตัดคำให้มีประสิทธิภาพด้วยเช่นกัน ในส่วนของระบบค้นหาข้อมูลข้ามภาษาไทยและอังกฤษนั้น Jaruskulchai (2002) ได้พัฒนาระบบค้นเอกสารข้ามภาษาไทยและภาษาอังกฤษขึ้น โดยมุ่งเน้นไปที่การแปลงคิวรี่ภาษาไทยเป็นภาษาอังกฤษ นอกจากนี้ Suwanvisat and Prasitjutrakul (2000) ได้พยายามปรับปรุงการเปรียบเทียบเอกสารข้ามภาษาโดยใช้หลักการของการออกเสียง (phonetic) เข้ามาช่วย ผลการทดลองพบว่าประสิทธิภาพการค้นคืนข้อมูลสูงถึง 80% แต่อย่างไรก็ตามระบบดังกล่าวยังไม่สามารถค้นหาเอกสารโดยใช้หลักการของการค้นหาเชิงความหมาย (semantic search) ได้และไม่สามารถแก้ปัญหาความกำกวมของคำศัพท์ภาษาไทยและอังกฤษได้อีกด้วย ดังนั้นสำหรับระบบค้นคืนเอกสารข้ามภาษาไทยและภาษาอังกฤษนั้น ยังจำเป็นต้องมีการพัฒนาอีกมากเพื่อแก้ปัญหาดังกล่าว เพราะระบบที่มีอยู่ในปัจจุบันยังไม่มีระบบได้มุ่งเน้นที่จะแก้ปัญหาเหล่านี้ ในบทความนี้จะขอยกตัวอย่างแนวทางการพัฒนา CLIR ในภาษาต่าง ๆ ในทวีปเอเชีย รวมถึงภาษาไทยและวิเคราะห์ถึงข้อจำกัดของเทคนิคต่างๆ ที่ถูกนำมาใช้ในแต่ละงานวิจัยเพื่อเป็นแนวทางสำหรับนักวิจัยไทยตลอดจนนักศึกษาในการพัฒนาระบบ CLIR ต่อไป



รูปที่ 1 ขั้นตอนการฝึกแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กให้เรียนรู้การสร้างรหัสคำ



รูปที่ 2 แนวคิดการการค้นคืนข้ามภาษาจีน–อังกฤษ

2.1 แนวคิดการค้นคืนข้อมูลข้ามภาษาไทย-อังกฤษ

ทัศนวรรณ (2543) ใช้เทคนิคการเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษโดยใช้นิวรอลเน็ตเวิร์ก (neural network) เพื่อการค้นคืนข้อมูลข้ามภาษา นิวรอลเน็ตเวิร์กที่ใช้เป็นแบบ แบ็กพรอพาเกชั่น (back propagation) ขั้นตอนการค้นคืนคำข้ามภาษาอาศัยการเปรียบเทียบรหัสของคำแบบประมาณ โดยอนุญาตให้มีความแตกต่างของรหัสที่นำมาเปรียบเทียบได้ไม่เกินหนึ่ง ผลการทดลองด้วยวิธี K–fold cross validation พบว่าการค้นคืนทั้งกรณีคำอังกฤษทับศัพท์คำไทยและกรณีคำไทยทับศัพท์อังกฤษมีประสิทธิภาพสูงเกิน 80% ขั้นตอนการทำงานของงานวิจัยนี้แสดงดังรูปที่ 1 ในส่วนของแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กจะแบ่งการทำงานออกเป็น 3 ชั้น (layer) และสามารถอธิบายได้ดังนี้

1) ชั้นอินพุต (input layer) ประกอบด้วยจำนวนนิวรอนเท่ากับจำนวนอักขระทั้งหมดของภาษาที่พิจารณา คูณด้วยจำนวนตัวอักษรทั้งหมดที่ใช้พิจารณา (กำหนดให้เป็นค่าคงตัว m) ดังนั้นกรณีคำอังกฤษ ข้อมูลเข้าจะมีจำนวน 26 × m นิวรอน และกรณีคำไทย ข้อมูลเข้าจะมีจำนวน 61 × m นิวรอน
2) ชั้นซ่อน (hidden layer) ได้ทำการทดลองเพื่อหาจำนวนนิวรอนที่เหมาะสม สำหรับแต่ละเน็ตเวิร์กซึ่งสำหรับคำไทยจะมี 61 นิวรอน คำอังกฤษจะมี 234 นิวรอน
3) ชั้นเอาท์พุต (output layer) จะมีจำนวนนิวรอนเท่ากับรหัสเสียงพยัญชนะและเสียงสระที่เป็นไปได้ทั้งหมด ซึ่งในกรณีคำไทยและคำอังกฤษทับศัพท์คำไทยจะมี 35 นิวรอน และในกรณีคำอังกฤษและคำไทยทับศัพท์คำอังกฤษจะมี 39 นิวรอน

ข้อจำกัดที่สำคัญของวิธีการนี้คือระบบยังไม่สามารถจัดการกับความกำกวมของภาษาไทยและภาษาอังกฤษได้หรือเรียกว่ายังไม่รองรับการค้นหาข้อมูลเชิงความหมาย (semantic search) นั่นเอง

2.2 แนวคิดการค้นคืนข้อมูลข้ามภาษาจีน-อังกฤษ

Tao Zhang et al. (2008) ได้คิดค้นเทคนิคในการแปลคิวรี่ภาษาจีนเป็นภาษาอังกฤษ (Chinese-English query translation) แสดงดังรูปที่ 2 โดยการใช้พจนานุกรมสองภาษา (bilingual dictionary) และใช้พจนานุกรมคำพ้อง (synonym) แต่ข้อเสียของวิธีการใช้พจนานุกรมคือไม่มีการพิจารณาความหมายของคำตามบริบท (context) ของประโยคซึ่งจะทำให้ได้ความหมายที่ผิดไปจากความหมายที่แท้จริงของคำในประโยคนั้น เนื่องจากคำศัพท์หนึ่งคำอาจมีหลายความหมาย ตัวอย่างเช่น “ผมเห็นว่าควรเลือกตั้งในเดือนมิถุนายน 2554” คำว่า “เห็น” ในบริบทนี้หมายถึง “มีความคิดว่า” ดังนั้นการใช้พจนานุกรมอาจจะแปลคำว่า “เห็น” นี่เป็นภาษาอังกฤษคำว่า

“look” หรือ “view” ซึ่งไม่ตรงกับความหมายที่แท้จริงเป็นต้น

2.3 แนวคิดการค้นคืนข้อมูลข้ามภาษาญี่ปุ่น-อังกฤษ

Yaoyong et al. (2006) ใช้วิธีการแปลโดยใช้พจนานุกรม 2 ภาษาเช่นกันและนำมาประยุกต์ใช้ร่วมกับแบบจำลองปริภูมิเวกเตอร์ในการคำนวณความคล้ายคลึง (similarity) ระหว่างคิวรี่และเอกสาร โดยใช้เทคนิคการวัดความเหมือนแบบโคไซน์ (cosine similarity) ดังนั้นก่อนที่จะแสดงผลลัพธ์ให้กับผู้ใช้ระบบค้นคืนสารสนเทศจะทำการคำนวณความคล้ายคลึงระหว่างเอกสารผลลัพธ์และคิวรี่ วิธีที่ใช้สำหรับการคำนวณความคล้ายคลึงระหว่างเอกสารและคิวรี่ ซึ่งเป็นวิธีการหาคล้ายคลึงกันจากค่าความต่างของมุมของข้อมูล 2 อันที่เกิดขึ้นบนปริภูมิเว็กเตอร์ ซึ่งความเหมือนแบบโคซายน์นี้จะมีค่าอยู่ระหว่าง 0-1 เท่านั้น วิธีการนี้เป็นที่นิยมและมีประสิทธิภาพสูงในการวัดความคล้ายคลึงระหว่างข้อมูล 2 อันและถูกนำมาประยุกต์ใช้กับศาสตร์การค้นคืนสารสนเทศอย่างแพร่หลาย วิธีการนี้จะมีประสิทธิภาพในกรณีที่เอกสาร 2 เอกสารมีความยาวไม่เท่ากัน

$$sim(d,q) = \frac{\sum_{i=1}^{n} d_i \times q}{(\sum_{i=1}^{n} d_i^2 \times \sum_{i=1}^{n} q^2)^{1/2}} \quad (1)$$

โดยที่ d คือเว็กเตอร์ของเอกสารต่าง ๆ ในคอลเล็คชั่น (collection) และ q คือเว็กเตอร์ของคิวรี่ที่ผู้ใช้ทำการใส่เข้าไป สรุปให้เข้าใจง่าย ๆ คือความคล้ายคลึงหาได้จากข้อมูลเอกสารและข้อมูลคิวรี่นั่นเอง

จากรูปที่ 3 สรุปการทำงานคือเมื่อผู้ใช้ใส่คิวรี่ซึ่งเป็นภาษาญี่ปุ่น (query in S) ระบบจะทำการแปลคิวรี่โดยใช้ตัวแปลภาษา (translator) ซึ่งอาศัยข้อมูลจากพจนานุกรมและการคำนวณทางสถิติ เมื่อได้ผลลัพธ์เป็นเอกสารต่างๆ (docs in T) และส่งกลับไปให้ผู้ใช้ (KWs in S) แต่ในกรณีที่เอกสารที่เป็นผลลัพธ์ไม่ใช้ภาษาญี่ปุ่น ระบบจะใช้การดึงคำสำคัญจากเอกสารและแปลเป็นภาษาญี่ปุ่นก่อนโดยใช้เครื่องแปลภาษา (machine translation-MT system)

จากการวิเคราะห์การใช้เทคนิคพจนานุกรมในระบบ CLIR จีน-อังกฤษและญี่ปุ่น-อังกฤษ นั้นพบว่าการใช้พจนานุกรมในระบบ CLIR มีข้อจำกัดสำคัญๆ 4 ประการ คือ

1) ผู้ใช้งาน CLIR ต้องมีความชำนาญสูงเนื่องจากต้องใช้คำศัพท์ที่เฉพาะเท่านั้น ดังนั้นเทคนิคนี้จึงยากต่อคนใช้ทั่วไป ข้อจำกัดนี้เป็นข้อจำกัดที่มีปัญหามากที่สุดเนื่องจากผู้ใช้จะต้องรู้ว่าคำศัพท์ในพจนานุกรมว่ามีอะไรบ้างซึ่งเป็นไปได้ยากสำหรับผู้ใช้ ดังนั้นจึงมีนักวิจัยพยายามที่จะพัฒนาวิธีอื่นที่ลดข้อจำกัดดังกล่าว เช่น ให้ผู้ใช้สามารถใช้คำศัพท์ได้อย่างอิสระ วิธีการที่ถูกพัฒนาขึ้นมาได้แก่ วิธีใช้คอร์ปัสหรือคลังข้อมูล (corpus-based) และวิธีใช้ฐานความรู้ (knowledge-based) (Diekema and Oard, 1998)
2) มีจำนวนคำศัพท์คงที่ การเพิ่มเติมคำศัพท์ภายหลังทำได้ยากเนื่องจากแต่ละภาษาจะมีคำศัพท์ใหม่ ๆ เพิ่มขึ้นทุกปี
3) การดูแลรักษาข้อมูลดัชนี (index information) ของเอกสารทั้ง 2 ภาษามีต้นทุนที่สูงมากกว่าปกติ ดังนั้นจึงไม่เหมาะกับระบบที่มีขนาดใหญ่
4) ส่วนข้อจำกัดอันสุดท้ายคือมีระบบที่มีอยู่ยังไม่รองรับการค้นหาข้อมูลเชิงความหมาย หมายความว่าระบบดังกล่าวทำการค้นหาข้อมูลโดยการเปรียบเทียบจากตัวอักขระเป็นหลัก แต่ไม่มีกลไกพิจารณาหาความหมายที่แท้จริงของคิวรี่ ตัวอย่างเช่น ถ้าผู้ใช้ใส่คิวรี่โดยใช้คำว่า car ระบบไม่ควรจะมองหาเฉพาะเอกสารที่มีคำว่า car

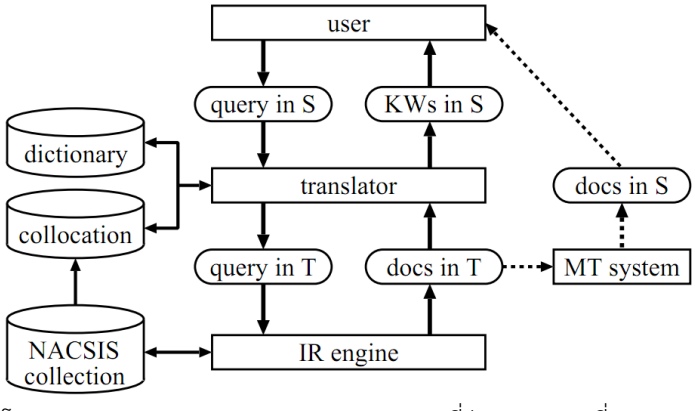
ปรากฏอยู่เท่านั้น แต่เอกสารใด ๆ ที่มีคำว่า BMW Honda และ Toyota หรือคำศัพท์อื่น ๆ ที่อยู่ในคอนเซฟท์ (concept) เดียวกัน คำศัพท์เหล่านี้ควรจะถูกพิจารณาว่าเป็นเอกสารที่เกี่ยวข้องกับคิวรี่เช่นกัน เนื่องจากทั้ง BMW Honda และ Toyota ถือเป็นคำที่มีความเกี่ยวข้องกับความหมายของคำว่า car ทั้งสิ้น ถึงแม้ว่าเอกสารเหล่านั้นจะไม่มีคำว่า car ปรากฏอยู่เลยก็ตาม ดังนั้นบทความนี้จะวิเคราะห์ถึงแนวคิดที่สำคัญสำหรับ ระบบค้นคืนสารสนเทศ เรียกว่า “การค้นหาข้อมูลเชิงความหมาย (semantic search)” ในบทความนี้จะอธิบายแนวคิดเกี่ยวกับการค้นหาเชิงความหมายในหัวข้อที่ 4

3. เทคนิคอื่น ๆ ที่นำมาใช้ในระบบ CLIR

เทคนิคในการพัฒนาระบบค้นคืนข้อมูลข้ามภาษานั้นสามารถทำได้หลายวิธี ในบทความนี้จะขอยกตัวอย่างเทคนิคสำคัญ ๆ ได้แก่ เทคนิคการแปลคิวรี่ (query translation) เทคนิคการแปลเอกสาร (document translation) และเทคนิคการสร้างข้อมูลที่เป็นตัวแทน (representation) ทั้งเอกสารและคำค้นหรือคิวรี่

3.1 การจับคู่กันระหว่างคิวรี่และดัชนี (matching)

ขั้นตอนสำคัญในการค้นหาเอกสารคือ ขั้นตอนการจับคู่ระหว่างคิวรี่และเอกสารที่ต้องการนี้เป็นขั้นตอนที่จะมีประสิทธิภาพเพียงใดขึ้นอยู่กับการทำดัชนี (index) เอกสารต่าง ๆ ในคอลเล็คชั่นและขั้นตอนการจับคู่ระหว่างคิวรี่และข้อมูลในดัชนีถือว่าเป็นองค์ประกอบที่สำคัญอันหนึ่งของระบบ CLIR จากปัญหาที่กล่าวมาข้างต้นคือปัญหาการใช้คำศัพท์ที่มีความกำกวม (มีหลายความหมาย) ทำให้ระบบค้นคืนข้อมูลทั้งแบบภาษาเดียวกันและข้ามภาษามีประสิทธิภาพที่ไม่ดีเท่าที่ควร ดังนั้นในหัวข้อนี้จะแนะนำเทคนิคที่นำมาใช้เพื่อลดปัญหาการจับคู่ระหว่างคิวรี่และข้อมูลดัชนี ซึ่งสามารถนำไปประยุกต์ใช้ได้ทั้งระบบค้น-คืนข้อมูลปกติและแบบข้ามภาษาได้



รูปที่ 3 การออกแบบโดยรวมของระบบ CLIR (S และ T แทนภาษาญี่ปุ่นและภาษาที่ถูกแปลความหมายตามลำดับ)

3.1.1 การแปลคิวรี่ (query translation)

เทคนิคการแปลคิวรี่คือ เทคนิคที่ทำการแปลคิวรี่ที่ผู้ใช้ใส่เข้ามาเป็นภาษาอื่น ๆ โดยอัตโนมัติ เทคนิคการแปลภาษานี้ถือเป็นเทคนิคที่มีประสิทธิภาพวิธีหนึ่งและเป็นเทคนิคที่สามารถทำได้แบบทันทีทันใด (real time) หลังจากผู้ใช้ใส่คิวรี่ ไม่จำเป็นต้องมีการคำนวณไว้ล่วงหน้า แต่ข้อจำกัดประการสำคัญของวิธีการนี้คือการแปลเป็นภาษาอื่น ๆ อาจจะแปลผิด เหตุผลสำคัญคือคิวรี่ที่ผู้ใช้ทำการใส่เข้ามาเป็นคิวรี่ที่สั้นเกินไปและไม่สามารถคาดเดาความหมายที่แท้จริงโดยดูจากบริบท (context) ได้ เนื่องจากคิวรี่ไม่ให้ข้อมูลในการวิเคราะห์ความหมายที่มากพอ คำศัพท์บางคำในคิวรี่มีความกำกวมและสามารถมีได้หลายความหมายจึงทำให้ระบบแปลความหมายของคิวรี่เป็นภาษาอื่น ๆ ที่ผิดเพี้ยนไป ส่งผลให้การค้นหาข้อมูลได้เอกสารที่ไม่เกี่ยวข้อง (non-relevance) สูง ดังนั้นในการพัฒนาระบบการแปลคิวรี่เป็นภาษาอื่น ควรคำนึงถึงปัญหาข้อนี้เป็นสำคัญ กรณีที่ผู้ใช้ทำการใส่คิวรี่ที่มีความยาวมากขึ้นและใช้เทคนิคของการวิเคราะห์ประโยคหรือวลีที่ดีพอจะสามารถช่วยลดข้อผิดพลาดที่เกิดขึ้นจากการแปลคิวรี่ได้

3.1.2 การแปลเอกสาร (document translation)

เทคนิคการแปลเอกสารคือเทคนิคที่ทำตรงกันข้ามกับเทคนิคการแปลคิวรี่ นั่นคือการแปลเอกสารทั้งหมดเป็นภาษาต่าง ๆ ที่ผู้ใช้อาจจะสร้างคิวรี่เป็นภาษานั้น ๆ ทั้งนี้วิธีการนี้อาจจะช่วยแก้ปัญหาของการแปลคิวรี่ เนื่องจากเอกสารประกอบด้วยข้อความที่ยาวกว่าคิวรี่ ดังนั้นระบบสามารถที่จะวิเคราะห์ความหมายของเอกสารโดยดูจากบริบทของคำต่าง ๆ ได้ดีกว่าคิวรี่และสามารถแก้ปัญหาความกำกวมของคำศัพท์ได้ แต่ปัญหาคือวิธีนี้ต้องใช้เวลาในการแปลมากกว่าคิวรี่ Gnter et al. (1997) ให้คำแนะนำว่าวิธีการแปลเอกสารนี้ควรใช้กับระบบที่มีเอกสารในคอเล็กชั่นไม่มากนักและเป็นเอกสารที่อยู่ในโดเมน (domain) ใดโดเมนหนึ่งเท่านั้น

3.1.3 เทคนิคการสร้างข้อมูลที่เป็นตัวแทนทั้งเอกสารและคิวรี่ (interlingual)

เทคนิคนี้เป็นเทคนิคที่ทำการแปลเอกสารและคิวรี่ให้อยู่ในรูปแบบที่มีโครงสร้างที่ใช้เป็นตัวแทนเอกสาร (representation) ที่ไม่ขึ้นกับภาษาใดภาษาหนึ่ง ตัวอย่างเช่นใช้เทคนิค LSI (Deerwester et al., 1990) เก็บข้อมูลเอกสารและคิวรี่ในรูปแบบของโมเดลเชิงพื้นที่เวกเตอร์หรือวิธี generalized vector space model ซึ่งทั้งสองวิธีจะทำการเรียนรู้ถึงความสัมพันธ์ระหว่างคำศัพท์ต่าง ๆ ในแต่ละภาษาจากการคำนวณทางสถิติและสามารถคำนวณหาความคล้ายคลึงระหว่างคำศัพท์ต่าง ๆ ทั้งที่อยู่ในภาษาเดียวกันและข้ามภาษาได้ หรืออาจจะเทคนิคของการสร้างฐานความรู้โดยใช้ออนโทโลจีในการเชื่อมโยงคำศัพท์ต่าง ๆ ข้ามภาษาก็ถือว่าเป็นเทคนิคที่กำลังได้รับความสนใจจากนักวิจัยต่าง ๆ ในปัจจุบัน จะเห็นว่าทั้ง 3 วิธีที่กล่าวมาข้างต้นนั้นมีความสัมพันธ์กับการแปลภาษาทั้งสิ้น ดังนั้นหัวใจสำคัญของระบบอีกประการหนึ่งนี้คือกลไกการแปลภาษา

3.2 เทคนิคในการแปลภาษา

ในบทความนี้มีจะแบ่งเทคนิคที่นำมาใช้ในการแปลภาษาซึ่งแบ่งได้เป็น 3 วิธีคือ 1) การใช้ออนโทโลจี 2) การใช้พจนานุกรม 2 ภาษา และ 3) การใช้เครื่องแปลเล็กซิคอน สำหรับวิธีพจนานุกรม 2 ภาษาได้อธิบายไปแล้วในหัวข้อ 2.2 ดังนั้นเนื้อหาในส่วนนี้จะอธิบายเฉพาะเรื่องการใช้ออนโทโลจีและการใช้เครื่องแปลเล็กซิคอน

3.2.1 ออนโทโลจี (ontology)

ออนโทโลจีเป็นโครงสร้างข้อมูลที่เก็บฐานความรู้ในลักษณะของลำดับชั้นของคอนเซฟท์และความสัมพันธ์ระหว่างคอนเซฟท์ต่าง ๆ ข้อดีของการนำเอาออนโทโลจีมาใช้ในระบบค้นคืนสารสนเทศข้ามภาษาคือสามารถรองรับการค้นหาเชิงความหมาย โดยที่ผู้ใช้ทั่วไปสามารถใช้คำ

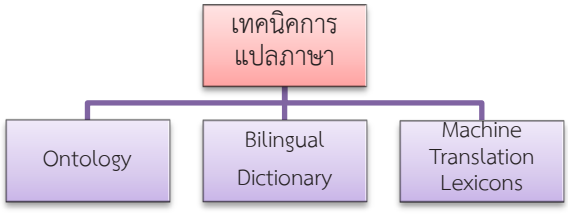
สำคัญในคิวรี่ได้อย่างอิสระ ระบบจะประมวลผลเพื่อหาคำที่มีความหมายเหมือนกัน หรือคำที่เกี่ยวข้องกัน (related terms) โดยดูจากคอนเซฟท์ (concept) ของคำสำคัญในคิวรี่เป็นหลัก ออนโทโลจีถือเป็นเทคนิคที่มีศักยภาพสูงในการประยุกต์ใช้กับระบบแปลภาษา โดยที่คำศัพท์ที่ต่างภาษากันจะอยู่ภายใต้คอนเซฟท์เดียวกัน แต่ปัญหาสำคัญคือการสร้างออนโทโลจียังอาจจะต้องสร้างด้วยมือ เนื่องจากยังไม่มีเทคนิคใดที่มีประสิทธิภาพมากพอในการเปลี่ยน (convert) จากข้อมูลในเอกสารเป็นคอนเซฟท์ต่าง ๆ ในออนโทโลจีโดยอัตโนมัติและยังมีอัลกอริทึมไม่มากนักในการหาคำศัพท์ที่สัมพันธ์กันแต่ถูกเขียนคนละภาษากันได้โดยอัตโนมัติ ดังนั้นการพัฒนาแนวคิดในการแก้ไขปัญหานี้จึงเป็นจุดหนึ่งที่น่าสนใจ

3.2.3 เครื่องแปลภาษาเล็คซิคอน (machine translation lexicon)

เทคนิคเครื่องแปลเลคซิคอนเป็นอีกวิธีหนึ่งที่นิยมนำมาใช้ช่วยในการแปลภาษาซึ่งจะทำการวิเคราะห์และทำการแปล วัตถุประสงค์สำคัญของเทคนิคเล็กซิคอนนี้คือการลดความกำกวมของคำศัพท์ต่าง ๆ และเลือกใช้ความหมายที่ถูกต้องเหมาะสมกับบริบท ในการนำเทคนิคเครื่องแปลเล็กซิคอนมาใช้กับระบบ CLIR นั้นทำได้ง่ายคือการนำเทคนิคนี้มาทำการแปลเพื่อหาความหมายของคิวรี่หรือความหมายของเอกสารนั่นเอง โดยทั่วไปแล้วระบบจะเลือกความหมายของคำศัพท์ที่เหมาะสมที่สุดมาเพียงความหมายเดียวเท่านั้น การที่ระบบเลือกเพียงความหมายเดียวนี้มีผลกระทบต่อประสิทธิภาพการค้นคืนข้อมูลของระบบเนื่องจาก ระบบค้นคืนสารสนเทศ อาจจะมองข้ามเอกสารที่เกี่ยวข้องบางเอกสารไป ตัวอย่างระบบแปลภาษาโดยใช้เครื่องแปลเล็กซิคอน เช่นในงานของ Oard et al. (1998)

4. แนวทางการพัฒนาระบบค้นคืนสารสนเทศเชิงความหมาย

การสร้างระบบค้นคืนสารสนเทศที่จะสามารถค้นหาข้อมูลเชิงความหมายได้นั้น จะต้องมีกลไกรองรับการค้นหาเชิงความหมาย ตัวอย่างเช่น การทำดัชนี การทำดัชนีแบบธรรมดา เช่น การทำดัชนีแบบอินเวิร์ท (inverted indexing) ไม่สามารถรองรับการค้นคืนข้อมูลเชิงความหมายได้ ดังนั้นจึงต้องมีการปรับเปลี่ยนวิธีการทำดัชนีที่จะสนับสนุนการค้นคืนข้อมูลเชิงความหมาย ในบทความนี้จะขอยกตัวอย่างการทำดัชนีเพื่อรองรับกลไกดังกล่าว 2 วิธี ได้แก่ วิธีการ latent semantic indexing (LSI) และวิธีการออนโทโลจี (ontology)



รูปที่ 4 เทคนิคการแปลข้อมูลเป็นภาษาต่าง ๆ

4.1 วิธี latent semantic indexing (LSI)

LSI เป็นวิธีการที่มีหลักการทำงานอยู่บนพื้นฐานของการคำนวณทางสถิติ โดยพิจารณาจากการปรากฏร่วมของคำต่าง ๆ (co-occurrence) เป็นเทคนิคการทำดัชนีแบบหนึ่งคิดโดย Deerwester et al. (1990) โดยใช้หลักการความสัมพันธ์ของคีย์เวิร์ดและเอกสารมาใช้ในการทำดัชนี นอกจากจะพิจารณาความถี่ของคำที่ปรากฏในเอกสารหนึ่งแล้ว ยังจะพิจารณาถึงความสัมพันธ์ของคำนั้น ๆ กับเอกสารอื่น ๆ ในคอลเล็คชั่นด้วย เรียกว่า "ความเกี่ยวข้องกัน (interrelationship)" LSI จะพิจารณาเอกสารใด ๆ ที่มีจำนวนคีย์เวิร์ดปรากกฏร่วมกันสูง จะถือว่าเอกสารเหล่านั้นมีความสัมพันธ์กันเชิงความหมาย (semantically related) เนื่องจากว่าเอกสารใด ๆ อาจจะมีความสัมพันธ์กันเชิงควมหมายถึงแม้ว่าเอกสารนั้น ๆ จะไม่มีคีย์เวิร์ดที่ซ้ำกันเลยก็ตาม แต่ LSI สามารถหาความสัมพันธ์ของเอกสารเหล่านั้นได้จากการคำนวณทางสถิติ ดังนั้น LSI จึงหาคำตอบของคิวรี่ที่มีประสิทธิภาพดีกว่าวิธีการหาเอกสารแบบเปรียบเทียบสายอักขระ (string matching) ถึงแม้ว่าเอกสารนั้น ๆ จะไม่ปรากฏคีย์เวิร์ดในคิวรี่เลยก็ตาม ดังนั้นด้วยวิธี LSI สามารถที่จะทราบได้ว่าเอกสารเกี่ยวกับ BMW Toyota และ Honda เป็นเอกสารที่ควรจะเป็นคำตอบของคิวรี่ Car แม้ว่าเอกสารเหล่านั้นจะไม่มีคำว่า Car ปรากฏอยู่เลยก็ตาม

วิธีการทำงานของ LSI โดยสรุปคือหลังจากที่ทำการเตรียมเอกสารและตัดคำที่ไม่สำคัญทิ้ง (stop words) เรียบร้อยแล้ว ขั้นแรกของวิธีการ LSI คือการสร้างเมตริกซ์ความสัมพันธ์ระหว่างคีย์เวิร์ดและเอกสาร (term-document matrix) เมตริกซ์นี้เป็นตารางที่เก็บความถี่ของคำต่าง ๆ ที่ปรากฏในเอกสารทั้งหมดในคอลเล็คชั่น หลังจากนั้นคำเอกสารในเมตริกซ์จะมีการให้น้ำหนักและทำการนอร์มอไลซ์เพื่อทำให้การเปรียบเทียบเอกสาร 2 เอกสารที่มีความยาวไม่เท่ากันมีความถูกต้องมากยิ่งขึ้นโดยใช้สมการที่ (2)

$$W_{ij} = L_{ij} G_i N_j \quad (2)$$

ซึ่ง W_{ij} คือน้ำหนักของแต่ละคำ L_{ij} คือน้ำหนักของคีย์เวิร์ด i ในเอกสาร j เรียกว่า "น้ำหนักแบบโลคอล (Local weight)" G_i คือน้ำหนักของคีย์เวิร์ด i ในคอเล็คชั่นเรียกว่า "น้ำหนักแบบโกบอล (Global weight)" และ N_j คือค่านอร์มอลไลซ์สำหรับเอกสาร j ในการคำนวณน้ำหนักแบบต่าง ๆ นั้นสามารถทำได้หลายรูปแบบ ผู้อ่านสามารถอ่านรายละเอียดเพิ่มเติมได้ใน (Chisholm and Kolda, 1999) อย่างไรก็ตามปัญหาที่เกิดขึ้นตามมาคือในกรณีที่มีเอกสารจำนวนมากทำให้เมตริกซ์ความสัมพันธ์ระหว่างคีย์เวิร์ดและเอกสารมีขนาดใหญ่มาก วิธีแก้ไขปัญหานี้คือการลดขนาดของเมตริกซ์ให้เล็กลงวิธีที่นิยมคือการใช้หลักการของ singular value decomposition (SVD) (Garcia, 2006)

4.2 การทำดัชนีโดยใช้ออนโทโลจี (ontology indexing)

เนื่องจากมนุษย์มีระบบความคิดและการรับรู้ถึงความหมายของสิ่งต่าง ๆ ได้ง่ายขึ้น ถ้าความสัมพันธ์ของสิ่งต่าง ๆ แสดงในรูปแบบของโครงสร้างลำดับชั้น (hierarchical model) ดังนั้นจึงเกิดแนวคิดการทำดัชนีในรูปแบบของโครงสร้างลำดับชั้น ซึ่งเป็นลักษณะคล้ายกลับโครงสร้างต้นไม้ (tree) ออนโทโลจี (ontology) คือการจัดองค์ความรู้หรือฐานความรู้ให้อยู่ในรูปแบบโคงสร้างลำดับชั้น โดยประกอบด้วยคอนเซฟท์ต่าง ๆ และความสัมพันธ์ของคอนเซฟท์ เรียกว่า "relationship" โครงสร้างนี้อำนวยความสะดวกให้กับกระบวนการค้นหาสารสนเทศเชิงความหมายให้มีประสิทธิภาพมากขึ้น ประโยชน์ของการทำดัชนีด้วย

ออนโทโลจีที่เหนือกว่าการทำดัชนีด้วยวิธีอื่น ๆ สามารถสรุปได้ดังต่อไปนี้

1) ความสัมพันธ์ของคอนเซฟท์ต่าง ๆ ในออนโทโลจีเป็นตัวกำหนดความหมายที่แท้จริงของคีย์เวิร์ดต่าง ๆ ซึ่งจะช่วยแก้ปัญหาความกำกวมของคีย์เวิร์ด เช่น หนึ่งคำมีหลายความหมาย (homonym) หรือคำหลาย ๆ คำมีความหมายเดียวกัน (synonym) (Chandrasekaran et al., 1999)
2) ออนโทโลจีใช้เป็นตัวกำหนดลำดับชั้นของคอนเซฟท์ต่าง ๆ (taxonomy) ภายในโดเมนที่สนใจ (Gasevic et al., 2009) ซึ่งแต่ละคอนเซฟท์จะจัดกลุ่มข้อมูลที่สัมพันธ์กันไว้ในกลุ่มเดียวกัน ซึ่งโครงสร้างลำดับชั้นนี้จะเป็นตัวช่วยในการค้นคืนเอกสารเชิงความหมายนั่นเอง
3) ออนโทโลจีสามารถที่จะถูกใช้ร่วมกันระหว่างแอปพลิเคชั่นหรือนำมาใช้ใหม่ได้ (Gasevic et al., 2009) เนื่องจากออนโทโลจีเป็นแหล่งคีย์เวิร์ดที่ใช้อธิบายคอนเซฟท์ต่าง ๆ ของโดเมน (domain) รูปแบบการเก็บข้อมูลดัชนีสามารถที่จะใช้ข้อมูลร่วมกันระหว่างแอปพลิคชั่นต่าง ๆ ได้โดยวิธีการต่อไปนี้ (Neches et al., 1991)

3.1) ใช้ข้อมูลร่วมกันโดยทำการสำเนาฐานความรู้ไปใช้ในแอปพลิเคชั่นใหม่โดยอาจจะมีการปรับปรุงโครงสร้างหรือทำการเพิ่มเติมองค์ความรู้ให้มากขึ้นและปรับใช้กับแอปพลิเคชั่นใหม่ได้

3.2) ใช้ข้อมูลร่วมกันผ่านทางซอฟต์แวร์เอเจนท์ ซึ่งสามารถเข้าถึงฐานความรู้ของออนโทโลจีได้

3.3) ใช้ข้อมูลร่วมกันผ่านทางเว็บเซอร์วิสเทคโนโลยี (web service)

จะเห็นว่าประโยชน์ของออนโทโลจีมีอยู่อย่างมากมายและสามารถนำมาประยุกต์ใช้กับระบบค้นคืนสารสนเทศได้อย่างมีประสิทธิภาพ แต่ข้อจำกัดสำคัญของออนโทโลจีคือไม่สามารถสร้างหรือออนโทโลจีให้ครอบคลุมข้อมูลในทุก ๆ ส่วนภายในโดเมนที่สนใจ หมายความว่าออนโทโลจีมีความไม่สมบูรณ์ของข้อมูล ดังนั้นนักวิจัยควรพึงระวังและคิดวิธีรองรับเมื่อระบบค้นคืนสารสนเทศหาข้อมูลที่ผู้ใช้ต้องการไม่พบ แนวทางการแก้ปัญหานี้แบบหนึ่งคือระบบควรพยายามใช้เทคนิคสำรองเช่น LSI ในการค้นหาข้อมูลที่ผู้ใช้ต้องการอีกครั้ง ในส่วนของการแทน (representation) ข้อมูลต่าง ๆ ในออนโทโลจีนี้จะเก็บข้อมูลได้หลายรูปแบบที่นิยมและรู้จักกันแพร่หลายได้แก่ RDF RDF(S) และ OWL ภาษาต่าง ๆ เหล่านี้มีข้อดีและข้อเสียต่าง ๆ กัน ซึ่งแล้วแต่ผู้พัฒนาระบบจะเลือกใช้ ในส่วนของการเข้าถึงข้อมูล (access) ในฐานความรู้จะมีลักษณะคล้าย ๆ กับเทคนิคการเข้าถึงข้อมูลแบบ RDBMS ซึ่งมีการทำงานอยู่บนพื้นฐานของ SQL (structure query language) และถูกดัดแปลงมาใช้เพื่อเป็นภาษาในการเข้าถึงข้อมูลในออนโทโลจีโดยเฉพาะ ตัวอย่างภาษาที่ใช้ในการเข้าถึงข้อมูลในออนโทโลจีได้แก่ SPARQL ตัวอย่าง SPARQL เพื่อค้นหาคนที่มีอายุมากกว่า 30 ปี ในฐานความรู้แสดงดังรูปที่ 5

```
PREFIX ns1: <http://www.w3.org/2001/vcard-rdf/3.0#>
      ns2: http://sampleVocabulary.org/1.3/People#
SELECT ?givenName, ?age
FROM  <Employee.rdf>
WHERE {?x ns1:N ?blank.
          ?blank ns1:Given ?givenName.
          ?x ns2:age ?age.
          FILTER (?age > 30)  }
```

รูปที่ 5 ตัวอย่าง SPARQL เพื่อค้นหาข้อมูลในออนโทโลจี

5. ความท้าทายในการพัฒนาระบบค้นคืนสารสนเทศข้ามภาษา

ในการพัฒนาระบบ CLIR นั้น มีสิ่งที่ผู้พัฒนาระบบต้องคำนึงถึงหลายประการ บทความนี้ขอสรุปสิ่งสำคัญที่ต้องคำนึงถึงในการพัฒนาระบบ CLIR ดังต่อไปนี้

1) ดัชนี (index): นักวิจัยต้องทำการออกแบบการทำดัชนีข้อมูลของเอกสารแต่ละภาษา เช่นเอกสารแต่ละภาษาควรทำดัชนีร่วมกันหรือควรจะทำดัชนีแยกกันแต่ละภาษาและจะเชื่อมโยงดัชนีเหล่านั้นได้อย่างไร การใช้ออนโทโลจีเป็นแนวทางการแก้ปัญหาหนึ่งของการทำดัชนีเพื่อรองรับข้อมูล 2 ภาษา
2) คิวรี่ (query): จะทำการลดความกำกวมของคำศัพท์ในคิวรี่อย่างไร คำศัพท์ในคิวรี่ควรจะขยายต่อ (expand) เพื่อหาคำที่เกี่ยวข้องหรือไม่และการขยายต่อนี้ควรทำเพียงภาษาเดียว เช่นภาษาเดียวกับคิวรี่หรือทำการขยายต่อทั้งสองภาษา
3) การเรียงลำดับผลลัพธ์ (ranking): การเรียงลำดับของเอกสารที่เป็นผลลัพธ์จะมีวิธีการเรียงลำดับข้อมูลอย่างไรหากผลลัพธ์ประกอบด้วยเอกสารมากกว่าหนึ่งภาษาและควรจะต้องเป็นการเรียงลำดับผลลัพธ์โดยพิจารณาความหมายเป็นหลักด้วยหรือเรียกว่า “semantic similarity”
4) ความเกี่ยวข้อง (relevance): หากมีการเลือกเอกสารที่เกี่ยวข้องโดยผู้ใช้ การคำนวณความเกี่ยวข้องระหว่างเอกสารที่ผู้ใช้เลือกและเอกสารอื่น ๆ ในคอลเล็คชั่นที่เขียนคนละภาษาจะคำนวณอย่างไร
5) การเปรียบเทียบคำในคิวรี่และคำในเอกสาร (matching): การเปรียบเทียบคำในคิวรี่และคำที่ทำดัชนีไว้นั้น ควรเป็นการเปรียบเทียบเชิงความหมาย ไม่ใช่การเปรียบเทียบโดยดูจากตัวอักขระเพียงอย่างเดียว

6. บทสรุป

จะเห็นว่าแนวคิดสำหรับวิธีการคืนคืนสารสนเทศข้ามภาษา (CLIR) คือการทำให้ระบบค้นหาข้อมูลทำงานได้อย่างเต็มประสิทธิภาพและเป็นสากล (universal) นักวิจัยในส่วนของการค้นคืนข้อมูลข้ามภาษาได้พยายามหาวิธีการที่จะรองรับกระบวนการที่จะค้นหาเอกสารที่ถูกเขียนขึ้นโดยใช้ภาษาที่แตกต่างไปจากภาษาในคิวรี่ ดังนั้นแนวคิดเกี่ยวกับการค้นคืนข้อมูลข้ามภาษาจึงมีประโยชน์กับคนทั่วโลก แนวความคิดนี้ถือเป็นแนวโน้มที่สำคัญสำหรับระบบค้นหาสารสนเทศและมีศักยภาพสูงในการพัฒนาต่อยอดในเชิงพาณิชย์เนื่องจากปัจจุบันผู้ใช้อินเทอร์เน็ตเกือบทุกคนมีการใช้ ระบบค้นหาข้อมูล (search

engine) ทุกวัน และชัดเจนว่าแนวคิดนี้ได้รับความสนใจจาก Google เช่นกันดังจะเห็นว่ามี Google เริ่มมีระบบช่วยแปลภาษาเป็นภาษาต่าง ๆ มาให้ และ Google จะใช้ข้อมูลที่ผู้ใช้ทำการแปลภาษานี้มาช่วยในการพัฒนาระบบ CLIR ในอนาคต

อย่างไรก็ตามความท้าทายที่สำคัญสำหรับนักวิจัยคนไทยในการพัฒนาระบบ CLIR คือจะออกแบบกลไกการทำดัชนีอย่างไรเพื่อรองรับการทำดัชนีที่มากกว่าหนึ่งภาษาและมีความถูกต้อง รวดเร็วในกาค้นหามากยิ่งขึ้น นอกจากนี้นักวิจัยไทยควรพัฒนากลไกการลดความกำกวมของคำในคิวรี่ทั้งนี้เนื่องจากถ้าหากระบบเข้าใจความหมายที่แท้จริงของคำค้นหรือคิวรี่จะสามารถเลือกเอกสารที่มีความสอดคล้องกันเชิงความหมายขึ้นมาแสดงต่อผู้ใช้มากยิ่งขึ้น และนอกจากนี้ยังต้องคำนึงถึงวิธีการเรียงลำดับผลลัพธ์ที่ได้จากการค้นหาอีกด้วยเพื่อให้ผู้ใช้คัดกรองเอกสารที่ไม่ต้องการออกไปได้ง่ายขึ้น

แนวทางหนึ่งที่นิยมในการแก้ไขปัญหาดังกล่าวคือการใช้ออนโทโลจีเพื่อรองรับการทำดัชนีมากกว่าหนึ่งภาษาและสามารถลดความกำกวมของคำต่าง ๆ โดยใช้โครงสร้างและความสัมพันธ์ของข้อมูลต่าง ๆ ในออนโทโลจีช่วยวิเคราะห์ความหมายที่แท้จริงของคำเหล่านั้นและเข้าใจความหมายของคิวรี่ได้อย่างมีประสิทธิภาพ หรือเรียกว่า การค้นหาสารสนเทศเชิงความหมาย “semantic-based information retrieval” นั่นเอง นักวิจัยหลายท่านได้นำแนวคิดนี้มาประยุกต์ใช้กับระบบค้นคืนสารสนเทศและพบว่าประสิทธิภาพการค้นคืน (precision และ recall) สูงขึ้น

7. เอกสารอ้างอิง

ทัศนวรรณ ศูนย์กลาง. (2543). การเข้ารหัสคำทับศัพท์ภาษาไทย/อังกฤษ เพื่อการค้นคืนข้ามภาษาด้วยเทคนิคนิวรอลเน็ตเวิร์ก (วิทยานิพนธ์มหาบัณฑิต). จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพมหานคร

Ampazis, N. and Iakovaki, H. (2004). Cross-language Information Retrieval using Latent Semantic Indexing and Self-Organizing Maps. Proceedings 2004 IEEE International Joint Conference on Neural Networks 1: 751–755.

Chandrasekaran, B., Josephson, J. R. and Benjamins, V. R. (1999). What Are Ontologies and Why Do We Need Them? IEEE Intelligent Systems 14(1): 20–26.

Chen, A. (2002). Multilingual Information Retrieval Using English and Chinese Queries. the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems pp. 44–58.

Chisholm, E. and Kolda, T. G. (1999). New Term Weighting Formulas For The Vector Space Method In Information Retrieval. USA: Computer Science and Mathematics Division, Oak Ridge National Laboratory.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41: 391–407.

Diekema, A. and Oard, D. (1998). Cross-Language Information Retrieval. Annual Review of Information Science 33.

Fredric C., G. and Douglas, W. O. (2001). The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries. the 10th Text Retrieval Conference 114–121.

Garcia, E. (2006). Singular Value Decomposition (SVD) A Fast Track Tutorial. http://www.miislita.com.

Gasevic, D., Djuric, D. and Devedzic, V. (2009). Model Driven Engineering and Ontology

Development (2nd ed.). London, United Kingdom: Springer.

Gnter, G. E., Erbach, G., Neumann, G. and Uszkoreit, H. (1997). Multilingual Indexing, Navigation and Editing Extensions for the World-Wide Web. Proceedings of the 3rd DELOS workshop - Cross-Language Information Retrieval 22–28.

Hull, D. A. and Grefenstette, G. (1996). Querying Across Languages: a Dictionary-based Approach to Multilingual Information Retrieval. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '96 49–57.

Jaruskulchai, C. (2002). Dictionary-Based Thai CLIR: An Experimental Survey of Thai CLIR. the 2nd Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems 209–218.

Li, Y. and Shawe-Taylor, J. (2006). Using KCCA for Japanese-English cross-language information retrieval and document classification. Journal of Intelligent Information Systems 27(2): 117–133.

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T. and Swartout, W. R. (1991). Enabling Technology for Knowledge Sharing. AI Magazine 12(3): 36–56.

Oard, D. (1997). Alignment of Spanish and English TREC Topic Descriptions. The Fifth Text Retrieval Conference 547–553.

Oard, D. W. and Hackett, P. (1998). Document Translation for Cross-Language Text Retrieval at the University of Maryland. The 6th Text Retrieval Conference (TREC-6). National Institutes of Standards and Technology 687–696.

Salton, G. (1969). Automatic Processing of Foreign Language Documents. 21: 1–28.

Salton, G. (1973). Experiments in Multi-Lingual Information Retrieval. Information Processing Letters 2(1): 6–11.

Suwanvisat, P. and Prasitjutrakul, S. (2000). Thai-English Cross-Language Transliterated Word Retrieval using Soundex Technique. National Computer Science and Engineering Conference, Bangkok. pp. 1-6.

Tao, Z. and Yue-Jie, Z. (2008). Research on Chinese-English Cross-Language Information Retrieval. International Conference on Machine Learning and Cybernetics 5: 2591–2596.

Yang, Y., Carbonell, J. G., Brown, R. D., & Frederking, R. E. (1998). Translingual Information Retrieval: Learning from Bilingual Corpora. Artificial Intelligence 103(1-2): 323–345.

❑❑❑❑❑