Abstract

Generating a valid metabolic pathway from biomedical research articles is a labor-intensive task for biologists. We developed an event extraction system, a text-mining task, to automatically discover metabolic interactions in research literature and then reconstruct metabolic pathways. The proposed system consists of the pipeline of four supervised-learning steps: named entity recognition, trigger detection, edge detection, and event reconstruction. For preparing the structured data used to train models, we need a powerful feature engineering method. As such, we proposed the Bag-of-Entities indexing method, which can transform text data into numerical vector effectively. The major obstacle in training a supervised model is a lack of a sufficient corpus of the domain of interest, because a corpus construction is a time-consuming task and needs specialist annotation. In particular, we introduced a multitask learning algorithm, a transfer-learning paradigm, which can leverage additional resources of an (existing) source domain to facilitate a classification of the metabolic event extraction system in the (focusing) target domain. To demonstrate a proof-of-concept of multitask learning classification, edge detection, the core step in our event extraction system, was used as a case study. The experimental results showed that the proposed event extraction system provided competitive performance against those of state-of-the-art related systems. In addition, using the proposed indexing method outperformed traditional and state-of-the-art word-vector representation approaches in term of resulting accuracy in a specific classification problem. Ultimately, the proposed multitask-learning can improve the performance of edge detection benefiting the overall performance of the event extraction system.

บทคัดย่อ

การสังเคราะห์วิถีเมแทบอลิซึมหรือชุดของปฏิกิริยาที่ใช้เอนไซม์ จากบทความวิจัยทางชีวการแพทย์นับว่าเป็นงานที่ยากสำหรับนักชีววิทยา เราจึงพัฒนาระบบการทำเหมืองข้อความเพื่อสกัดเหตุการณ์หรือปฏิสัมพันธ์ระหว่างเมแทบอลิซึมที่เกิดขึ้นในเซลล์ จากบทความอิเล็กทรอนิกส์ โดยอัตโนมัติ เหตุการณ์ที่ได้ถูกนำไปสกัดวิถีเมแทบอลิซึม ระบบที่พัฒนาประกอบด้วย 4 ขั้นตอนสำคัญ ได้แก่ การตรวจจับตำแหน่งของคำและจำแนกชนิดของคำ การตรวจจับคำที่แสดงถึงการเกิดขึ้นของเหตุการณ์ การจัดกลุ่มหน้าที่ของคำในประโยค และการสร้างเหตุการณ์ขึ้นใหม่ เนื่องจากการฝึกสอนโมเดลอัตโนมัติสำหรับแต่ละงาน จำเป็นต้องมีคุณลักษณะของชุดฝึกสอนที่มีคุณภาพ เราจึงนำเสนอวิธีการจัดทำดัชนีของกลุ่มคำ ซึ่งสามารถใช้แปลงข้อมูลที่เป็นข้อความให้อยู่ในรูปตัวเลขได้อย่างมีประสิทธิภาพ นอกจากนั้นอุปสรรคสำคัญในการจัดเตรียมชุดฝึกสอนโมเดล คือความสมบูรณ์ครบถ้วนของชุดข้อมูลฝึกสอน เพราะการจัดทำชุดฝึกสอนเป็นงานที่ใช้เวลามากและต้องถูกจัดทำโดยผู้เชี่ยวชาญในสาขาวิชานั้นๆ ในงานวิจัยนี้เราจึงนำเสนออัลกอริทึมการเรียนรู้แบบหลายงานโดยอาศัยทฤษฎีการถ่ายโยงการเรียนรู้ ซึ่งสามารถใช้ประโยชน์จากทรัพยากรที่มีอยู่เดิม ในขอบเขตการเรียนรู้อื่น เพื่ออำนวยความสะดวกในการทำเหมืองข้อความในขอบเขตการเรียนรู้ทางชีวการแพทย์ที่สนใจ เพื่อแสดงให้เห็นถึงประสิทธิภาพของอัลกอริทึมที่นำเสนอ เราใช้การจัดกลุ่มหน้าที่ของคำในประโยคซึ่งเป็นขั้นตอนสำคัญในระบบการทำเหมืองข้อความที่นำเสนอเป็นกรณีศึกษา ผลการทดลองแสดงให้เห็นว่าระบบการทำเหมืองข้อความที่นำเสนอมีประสิทธิภาพดีเทียบเท่าหรือดีกว่าระบบที่มีในปัจจุบัน การจัดทำดัชนีที่นำเสนอสามารถจัดเตรียมข้อมูลที่ไม่เป็นโครงสร้าง ได้อย่างมีประสิทธิภาพ ดีกว่าการทำดัชนีที่มีในปัจจุบัน โดยส่งผลให้ความแม่นยำของโมเดลของปัญหาการจำแนกสูงขึ้น โดยเฉพาะอย่างยิ่งอัลกอริทึมการเรียนรู้แบบหลายงานที่นำเสนอสามารถเพิ่มประสิทธิภาพการจัดกลุ่มหน้าที่ของคำในประโยค ส่งผลโดยตรงต่อประสิทธิภาพโดยรวมของระบบการทำเหมืองข้อความ