การศึกษาเปรียบเทียบวิธีการตัดคำสำหรับการจำแนกประเภทข้อความในภาษาไทย

A Comparative Study of Word Segmentation on Thai Text Categorization

พัชรินทร์ ทองอารีย์, รัชฎา คงคะจันทร์

บทคัดย่อ

งานวิจัยนี้เป็นการศึกษาเชิงเปรียบเทียบเพื่อหาโมเดลที่มีประสิทธิภาพในการประยุกต์ใช้งานด้านการจำแนกประเภทข้อความภาษาไทย (Thai Text Categorization) ขั้นตอนที่มีผลโดยตรงต่อการสร้างโมเดลคือการเตรียมหน่วยคำได้อย่างถูกต้อง ในการเตรียมคำสำหรับข้อความภาษาไทยสามารถทำได้โดยอาศัยเทคนิคการตัดคำ (Word Segmentation) ที่ผ่านมามีการนำเสนออัลกอริทึมสำหรับการตัดคำจากข้อความภาษาไทยมากมายหลายวิธี แต่ยังไม่มีวิธีใดที่สามารถให้ผลความถูกต้องอย่างสมบูรณ์แบบ จุดมุ่งหมายหลักในงานวิจัยนี้คือการเปรียบเทียบหาวิธีการตัดคำที่มีประสิทธิภาพในการเตรียมชุดตัวแปรคำ (Term Features) สำหรับการสร้างโมเดลจำแนกประเภทข้อความภาษาไทยนอกจากนี้ในงานวิจัยยังมีการศึกษาเปรียบเทียบวิธีการให้ค่าถ่วงน้ำหนักมิติข้อมูล (Feature Weighting) แบบต่างๆ ได้แก่ไบนารี่ (Binary) ความถี่ของคำ (TF - Term Frequency!) ความถี่ของคำคูณส่วนกลับของความถี่ของเอกสาร(TFIDF - Term Frequency-Inverse Document Frequency) และค่าล็อกของความถี่ของคำคูณส่วนกลับของความถี่ของเอกสาร (LOG-TFIDF) จากผลการทดลองบนคลังเอกสารข่าวพบว่าโมเดลที่ประกอบไปด้วยวิธีการตัดคำโดยใช้เทคนิคการเรียนรู้ของเครื่องแบบConditional Random Field (CRF) พร้อมด้วยการให้ค่าถ่วงน้ำหนักมิติข้อมูลแบบ LOG-TFIDF มีประสิทธิภาพสูงที่สุดกับวิธีการจำแนกประเภทข้อมูลแบบ Support Vector Machine (SVM) โดผลลัพธ์ของ F1 มีค่าเท่ากับ 0.959

Abstract

In this paper, we performed a comparative study of different feature processing approaches for Thai text categorization. One of the most important steps which directly affect the quality of the classification model is the preparation of term features. For Thai texts, term features can be extracted by using word segmentation technique. Many Thai word segmentation algorithms have previously been proposed, however, none yielded perfect results. The main goal of this paper is to compare and find the most suitable word segmentation algorithm for preparing the term features for Thai text categorization. In addition, we also compared among different feature weighting strategies including Binary, Term Frequency (TF), Term Frequency-Inverse Document Frequency (TFIDF), and Log of TFIDF. The experiments on Thai news corpus showed that the best performance was obtained when the word segmentation model based on the Conditional Random Fields (CRFs) was applied with the term weighting approach of LOG-TFIDF. This combination of feature processing, when applied under the Support Vector Machines (SVMs) algorithm, achieved the highest F1 measure of 0.959.