ระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่าน
โดยใช้เครือข่ายประสาทเทียม

PASSAGE GRADING SYSTEM USING NEURAL NETWORK

วสันต์ ณ ชัย

WASAN NA CHAI

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2555

KMITL-2012-SC-M-002-006

PASSAGE GRADING SYSTEM USING NEURAL NETWORK

WASAN NA CHAI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2012

KMITL-2012-SC-M-002-006

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์	ระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านโดยใช้เครือข่ายประสาทเทียม Passage Grading System Using Neural Network
นักศึกษา	นายวสันต์ ณ ชัย
รหัสประจำตัว	50067506
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร.นวลสวาท หิรัญสกลวงศ์

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ผศ.ดร.กรกช	ประชุมรักษ์	
รศ.ดร.จีรพร	วีระพันธุ์	
ดร.ชาคริต	วัชโรภาส	
ผศ.ดร.นวลสวาท	หิรัญสกลวงศ์	

วัน / เดือน/ ปี ที่สอบ 10 เมษายน พ.ศ. 2555 เวลา 15.00 – 18.00 น.
สถานที่สอบ ณ ห้อง 216 ชั้น 1 อาคารจุฬาภรณวลัยลักษณ์ 1

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 18 **เดือน** เม.ย. **พ.ศ.** 55

หัวข้อวิทยานิพนธ์	ระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านโดยใช้เครือข่ายประสาทเทียม
นักศึกษา	นายวสันต์ ณ ชัย
รหัสประจำตัว	50067506
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2555
อาจารย์ที่ปรึกษา	ผศ.ดร.นวลสวาท หิรัญสกลวงศ์

บทคัดย่อ

ในการศึกษานี้ผู้วิจัยนำเสนองานวิจัยและพัฒนาเรื่องระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านโดยใช้เครือข่ายประสาทเทียม ระบบดังกล่าวจะเรียนรู้และสร้างโมเดลทางสถิติสำหรับแต่ละระดับชั้นการศึกษาของเนื้อความจากตัวอย่างภายในคลังเนื้อความ (passage corpus) ในการหาความแตกต่างระหว่างระดับชั้น มีการแยกองค์ประกอบออกเป็น 3 ส่วน คือ ความยากของคำ ความยาวของพยางค์ต่อคำ และความซับซ้อนของประโยค หลังจากนั้นคะแนนทั้ง 3 ส่วนจะถูกนำมาใช้ร่วมกันเพื่อสร้างเป็นโมเดลทางสถิติสำหรับแต่ละระดับชั้นการศึกษาโดยประยุกต์โครงข่ายประสาทเทียม (Neural Network) เพื่อปรับสมดุลค่าของคะแนน คลังเนื้อความที่ใช้ในการศึกษาชิ้นนี้ถูกเก็บรวบรวมมาจากเนื้อความภาษาอังกฤษสำหรับอ่านที่อยู่ในแบบเรียนที่กระทรวงศึกษาธิการประเทศไทยเป็นผู้อนุมัติ

งานชิ้นนี้ได้ให้คุณประโยชน์ในการนำเสนอแนวทางใหม่สำหรับระบบแบ่งแยกระดับเนื้อความภาษาอังกฤษสำหรับอ่านที่แต่เดิมจะใช้การกำหนดกฎเกณฑ์คะแนนของแต่ละระดับชั้นโดยนักภาษาศาสตร์หรืออาจารย์ ซึ่งกฎเกณฑ์นั้นจะไม่เหมาะสมในการนำมาประยุกต์ใช้กับมาตรฐานความรู้หรือการเรียนการสอนภาษาอังกฤษของแต่ละประเทศ นอกจากนั้นระบบดังกล่าวสามารถนำไปประยุกต์ใช้ร่วมกับการเรียนรู้ด้วยสื่ออิเล็กทรอนิกส์สำหรับการแบ่งแยกระดับความยากของเอกสารหรือเนื้อความภาษาอังกฤษในการเรียนรู้ได้อย่างอัตโนมัติ และยังเปิดโอกาสให้ผู้เรียนสามารถเลือกเนื้อความที่ตนเองสนใจได้อย่างอิสระ ระบบนี้ยังสามารถนำไปประยุกต์กับระบบสืบค้นข้อมูล (Information Retriever) เพื่อช่วยคัดกรองระดับของเนื้อหาของเว็บไซต์เพื่อให้ผู้ใช้ได้รู้ระดับของเนื้อหาและกำหนดความเหมาะสมโดยผู้ปกครองได้อีกด้วย

คำสำคัญ: ระบบวิเคราะห์การให้คะแนน, เนื้อความภาษาอังกฤษสำหรับอ่าน, ความสามารถและประสิทธิภาพในการอ่าน, การเรียนรู้แบบมีผู้สอน

Thesis Title	Passage Grading System Using Neural Network
Student	Wasan Na Chai
Student ID	50067506
Degree	Master of Science
Program	Computer Science
Year	2012
Thesis Advisor	Asst. Prof. Dr. Nualsawat Hiransakolwong

ABSTRACT

In this study, an English reading passage classification system using a statistical approach is purposed. A statistical model of each reading passage level is learned from a corpus. The reading passage corpus is collected from the actual reading passages and supplementary readings in school with its assigned level. Three complexity features, which are word, syllable and sentence complexity, are calculated separately to reflect passage characteristics. For the word level, a list of words within a corpus and a word list approved by the Ministry of Education of Thailand are criteria for a level standard. An average syllable of words is taken into account for representing a complexity based on the use of long words. A sentence complexity is a key for a complication of a sentence in terms of iterative complex sentences and compound sentences in the passage. These three feature scores are combined and tuned up with neural network for a level characteristic model of each level. With the learned models, the system compares a reading passage input and returns a probable level categorized by Thai academic school level.

The contribution of the purposed system is that it does not require manual criteria which can particularly be varied by either the differentiated English standard of each country or a personal prejudice within manual criteria, but it automatically classifies an English reading passage by comparing to the supervised model. The benefit of the system is to be applied in e-learning since it can reduce time consumption and burdens of teachers to manually select an appropriate passage to match students' interest. Moreover, information retrieval system can be plugged in to classify a level of the content in the website for self-improvement learning and a parental control.

Keywords: grading system, English reading passage, readability, supervised learning

กิตติกรรมประกาศ

วิทยานิพนธ์นี้มิอาจจะสำเร็จลุล่วงไปได้ด้วยดี หากมิได้รับคำแนะนำ คำชี้แจง ความรู้และความเอาใจใส่จาก ผศ.ดร.นวลสวาท หิรัญสกลวงศ์ ผู้เป็นอาจารย์ที่ปรึกษา ซึ่งท่านได้สละเวลาให้กับข้าพเจ้าอย่างเต็มที่ จึงใคร่ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ผศ.ดร.กรกช ประชุมรักษ์ ผศ.ดร.จีรพร วีระพันธุ์ และดร.ชาคริต วัชโรภาส คณะกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำตลอดจนถึงข้อชี้แนะและในที่สุดทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

นอกจากนี้ข้าพเจ้าขอขอบคุณ ดร.เทพชัย ทรัพย์นิธิ และ คุณธเนศ เรืองรจิตรปกรณ์ นักวิจัย ห้องปฏิบัติการวิจัยเทคโนโลยีภาษาธรรมชาติและความหมาย หน่วยวิจัยวิทยาการสารสนเทศ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (เนคเทค) สำหรับคำชี้แนะแนวทางและปัญหาต่างๆในการทำวิทยานิพนธ์เล่มนี้ และขอขอบคุณคุณณัฐพล กฤษสุทธิกุล สำหรับเครื่องมือในการทำวิจัย

สำหรับคุณงามความดีและประโยชน์อันใดที่เกิดขึ้นจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดา มารดา อาจารย์ทุกท่านซึ่งเป็นที่เคารพรักยิ่ง ตลอดจนญาติพี่น้อง และเพื่อนๆทุกคน

สุดท้ายนี้ขอขอบพระคุณบิดา มารดา ที่สนับสนุนและเป็นกำลังใจในระหว่างการศึกษาเป็นอย่างดียิ่งรวมทั้งอำนวยความสะดวกในทุกๆด้านพร้อมเป็นกำลังใจในการทำวิทยานิพนธ์อย่างดียิ่งตลอดมา

วสันต์ ณ ชัย

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการเรียนรู้ภาษาอังกฤษมีความสำคัญอย่างยิ่ง เนื่องจากเป็นภาษาสากลที่ใช้ในการติดต่อสื่อสารระหว่างประเทศ รวมถึงเป็นพื้นฐานในการศึกษาและเรียนรู้วิชาแขนงต่างๆ จึงเป็นเรื่องยากสำหรับการเรียนการสอนของประชากรที่ไม่ได้ใช้ภาษาอังกฤษเป็นภาษาหลัก อย่างเช่นประเทศไทย ให้มีประสิทธิภาพ เนื่องด้วยความแตกต่างของตระกูลภาษา ที่มีการใช้กฎและไวยากรณ์ ต่างๆ ที่ไม่คล้ายคลึงกับภาษาไทย จากค่าสถิติภายใน ปี 2552 ที่ทำการทดสอบวัดความรู้ภาษาอังกฤษสำหรับชาวต่างชาติ (Test of English as a Foreign Language: TOEFL)[1] ประเทศไทยมีคะแนนเฉลี่ยอยู่ที่ 493 คะแนน โดยต่ำกว่าคะแนนเฉลี่ยของทวีปเอเชีย คือ 514.6 คะแนน ดังนั้นประชากรในประเทศไทยยังมีความจำเป็นต้องฝึกฝนและพัฒนาทักษะภาษาอังกฤษเพิ่มขึ้นอีก

สำหรับทักษะภาษาอังกฤษนั้น สามารถแบ่งออกได้เป็น ฟัง พูด อ่าน เขียน การมุ่งเน้นพัฒนาทักษะด้านการอ่านมีความสำคัญมากเป็นอันดับต้นๆ เนื่องจากทักษะด้านการอ่านเป็นพื้นฐานไปสู่การพัฒนาทักษะด้านอื่นๆ ตลอดจนช่วยเรียนรู้ทั้งการใช้คำศัพท์และโครงสร้างไวยากรณ์ใหม่ๆ ได้ อย่างไรก็ตามการพัฒนาทักษะด้านการอ่านจะไม่สามารถพัฒนาขึ้นได้หากผู้เรียนไม่มีแรงจูงใจที่จะอ่านเนื้อความนั้น เนื่องจากเนื้อหาภายในเนื้อความนั้นไม่ตรงกับความสนใจของนักเรียน ซึ่งปัญหานี้เกิดจากเนื้อความส่วนมากที่ใช้ในการเรียนรู้และพัฒนาทั้งหมด ล้วนได้รับการกำหนดจากผู้สอนนั่นเอง โดยที่ผู้สอนหรืออาจารย์เป็นคัดเลือกมาถ่ายทอดให้กับนักเรียนตามรูปแบบไวยากรณ์หรือกลุ่มคำศัพท์ตามต้องการ ดังนั้นการอนุญาตให้นักเรียนหาบทความหรือเนื้อความภาษาอังกฤษที่ตนเองสนใจมาใช้ในการอ่านในห้องเรียนจึงเป็นแนวทางที่สำคัญอีกประการหนึ่งที่จะสามารถช่วยพัฒนาทักษะด้านการอ่านของนักเรียนให้เพิ่มขึ้นได้

อย่างไรก็ตามถ้าหากผู้สอนเลือกแนวทางในการพัฒนาทักษะการอ่านของนักเรียนโดยการอนุญาตให้นักเรียนสามารถเลือกอ่านบทความหรือเนื้อความที่ตนเองสนใจแล้ว ผู้สอนจำเป็นต้องดูแลและพิจารณาเนื้อความที่ถูกเลือกโดยตัวนักเรียนเองในมุมมองของความเหมาะสมในเรื่องความยากหรือง่ายให้ตรงกับระดับความยากที่ตนเองศึกษาอยู่ หากเนื้อความยากเกินไปนักเรียนจะต้องใช้เวลานานในการเข้าใจเนื้อหาและคำศัพท์ต่างๆ รวมทั้งรูปแบบทางไวยากรณ์ที่อยู่นอกเหนือจากที่เคยเรียน หรือถ้าหากเนื้อความง่ายเกินไปนักเรียนก็จะไม่สามารถพัฒนาเพิ่มพูนทักษะด้านการอ่านของตนเองได้ อีกทั้งยังทำให้เกิดความแตกต่างของความยากและความง่ายของเนื้อความในระดับชั้นเดียวกันได้ หากผู้สอนไม่ควบคุมมาตรฐานของเนื้อความที่นักเรียนนำมา ซึ่ง

ถ้าหากผู้สอนมีจำนวนนักเรียนภายใต้การดูแลมากก็จะทำให้มีภาระในการควบคุมความเหมาะสมของเนื้อความมากขึ้นไปด้วย ซึ่งงานในการคัดเลือกความเหมาะสมของเนื้อความนั้น เป็นงานที่เป็นภาระหนักแก่ผู้สอนอย่างมาก

ที่ผ่านมาได้มีการพัฒนา ระบบที่ช่วยคัดแบ่งระดับความยากของเนื้อความภาษาอังกฤษสำหรับอ่าน (English reading passage) อยู่หลายระบบ แต่ทั้งหมดล้วนเป็นระบบที่อิงอยู่กับการใช้กฎการให้คะแนน (heuristic rules) ด้วยมาตรฐานของการเรียนรู้ภาษาอังกฤษของเจ้าของภาษา (native English) ทำให้กฎเหล่านั้น ไม่สอดคล้องกับมาตรฐานการศึกษาของประเทศที่ใช้ภาษาอังกฤษเป็นภาษาที่สอง (English as second language: ESL) หรือ ภาษาอังกฤษเป็นภาษาต่างประเทศ (English as foreigner language: EFL) นอกจากนั้น มาตรฐานของแต่ละประเทศก็แตกต่างกัน ตามความเข้มข้นของการเรียนการสอนภาษาอังกฤษของแต่ละประเทศ ดังนั้น การพัฒนากฎการให้คะแนนหรือแบ่งระดับความยากจึงเป็นไปไม่ได้ในการประยุกต์ใช้ร่วมกันและเป็นสากล

การจะพัฒนาระบบที่ช่วยคัดแบ่งระดับความยากของเนื้อความสำหรับอ่าน (Passage Grading System) ที่สามารถใช้งานได้โดยทั่วไปนั้น การเรียนรู้ของระบบ (Machine learning) เป็นแนวทางหนึ่งที่จะช่วยสร้างการคัดแบ่งระดับความยากได้โดยอัตโนมัติ ซึ่งการคัดแบ่งจะขึ้นอยู่กับชุดข้อมูล (data set) ที่ระบบเรียนรู้ ดังนั้น หากเปลี่ยนชุดข้อมูลเป็นฐานข้อมูลเนื้อความสำหรับอ่านภาษาอังกฤษที่ใช้กันอยู่ภายในประเทศใด ระบบก็จะสามารถคัดแบ่งระดับความยากตามระดับที่เป็นมาตรฐานของประเทศนั้นได้ อย่างไรก็ตามการเรียนรู้ของระบบโดยการใช้ค่าทางสถิติอย่างเดียวมีผลงานวิจัยชี้ชัดมาแล้วว่า ความถูกต้องนั้นจะต่ำ ดังนั้น การใช้ความรู้ทางด้านภาษามาเป็นองค์ประกอบด้วยจะช่วยเพิ่มความถูกต้อง รวมทั้งการคัดแบ่งระดับความยากจะมีความสมเหตุสมผลและสอดคล้องตามหลักไวยากรณ์ภาษาที่อ้างอิงได้

งานวิจัยนี้นำเสนอการพัฒนาระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่าน (English reading passage) โดยประยุกต์กับแนวทางค่าสถิติและอิงกฎไวยากรณ์ภาษา ระบบดังกล่าวจะเรียนรู้และสร้างโมเดลทางสถิติสำหรับ ความยากของเนื้อความแต่ละระดับชั้นการศึกษาจากตัวอย่างภายในคลังเนื้อความ (passage corpus) ในการหาความแตกต่างระหว่างระดับชั้น มีการแยกปัจจัยตามหลักภาษาเพื่อช่วยคำนวณออกเป็น 3 ส่วน คือ ความยากของคำ ความยาวของพยางค์ต่อคำ และความซับซ้อนของประโยค หลังจากนั้นคะแนนทั้ง 3 ส่วนจะถูกนำมาใช้ร่วมกันเพื่อสร้างเป็นโมเดลทางสถิติสำหรับแต่ละระดับชั้นการศึกษาโดยประยุกต์โครงข่ายประสาทเทียม (Neural Network) เพื่อปรับสมดุลค่าของคะแนน ซึ่งการพัฒนาระบบที่นำเสนอนี้ จะสามารถช่วยลดภาระของผู้สอนในการวิเคราะห์ความเหมาะสมของเนื้อความที่นักเรียนนำเสนอ อีกทั้งนักเรียนสามารถนำระบบดังกล่าวมาตรวจสอบเนื้อความที่ตนเองสนใจว่าเหมาะสมกับระดับการศึกษาหรือไม่

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

วิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อพัฒนาระบบแบ่งแยกระดับเนื้อความภาษาอังกฤษสำหรับอ่านให้สามารถวิเคราะห์เนื้อความภาษาอังกฤษให้มีความเหมาะสมกับนักเรียนแต่ละระดับชั้นเรียนได้ ศึกษาอัลกอริทึมสำหรับการคำนวณคะแนนในการแบ่งแยกระดับของเนื้อความภาษาอังกฤษให้มีความเหมาะสมกับนักเรียนไทย โดยสมการที่นำมาคำนวนคะแนนจะต้องคิดจากความซับซ้อนในการเรียนรู้ภาษาอังกฤษตามหลักสูตรที่มีในห้องเรียน

1.3 สมมติฐานของการศึกษา

1) เนื้อความภาษาอังกฤษสำหรับอ่านสามารถนำมาคัดแบ่งระดับความยากได้ โดยใช้แง่มุมทางภาษาศาสตร์ได้แก่ จำนวนพยางค์ ความยากของคำศัพท์ และ ความซับซ้อนของประโยค

2) แง่มุมทางภาษาศาสตร์ทั้ง 3 สามารถพัฒนาเป็นสมการเพื่อมาคำนวณค่าระดับความยากได้

3) เครือข่ายประสาทเทียมสามารถนำมาใช้สร้างโมเดลความยากของเนื้อความตามระดับชั้นเรียนของนักเรียนไทย จากค่าระดับความยากทั้ง 3 จากทุกแง่มุมทางภาษาศาสตร์

4) โมเดลที่ได้จากการเรียนรู้ของเครือข่ายประสาทเทียมสามารถนำไปใช้คัดแบ่งระดับความยากของเนื้อความภาษาอังกฤษสำหรับอ่านแบ่งตามระดับชั้นตามหลักสูตรให้กับนักเรียนไทยได้อย่างเหมาะสม

1.4 ขอบเขตการวิจัย

วิทยานิพนธ์นี้มีขอบเขตการวิจัยเพื่อทำการศึกษาและคิดค้นอัลกอริทึมเพื่อวัดระดับความยากของเนื้อความภาษาอังกฤษสำหรับอ่าน โดยมุ่งเน้นไปที่เนื้อความของนักเรียนไทยในระดับชั้นประถมศึกษาชั้นปีที่ 1 ถึง 6 และ มัธยมศึกษาชั้นปีที่ 1 ถึง 6

อัลกอริทึมนี้จะประยุกต์ใช้แง่มุมทางภาษาศาสตร์เพื่อบ่งบอกระดับความยากที่แตกต่างกันของเนื้อความในแต่ละระดับชั้น และใช้การเรียนรู้ด้วยเครื่องแบบเครือข่ายประสาทเทียมเพื่อสร้างเป็นโมเดลทางสถิติของระดับความยากสำหรับนำมาใช้คัดแบ่งความยากของเนื้อความถัดไป กล่าวคือ อัลกอริทึมนี้เป็นการประยุกต์ระหว่างแนวทางการใช้กฎทางภาษาศาสตร์และการใช้ค่าสถิติ

ขอบเขตของการพัฒนาอัลกอริทึม คือ ใช้แง่มุมทางภาษาศาสตร์ 3 ชนิดได้แก่จำนวนพยางค์ ความยากของคำศัพท์ และ ความซับซ้อนของประโยค และใช้การเรียนรู้ด้วยเครื่องแบบเครือข่ายประสาทเทียมเพื่อสร้างโมเดลทางสถิติของระดับความยาก โดยจะใช้ฐานข้อมูลเนื้อความแบ่งเป็น 4 ช่วงระดับชั้น คือ ประถมต้น ประถมปลาย มัธยมต้น และมัธยมปลาย และมีเนื้อความต่อระดับชั้นละ 200 เนื้อความ

1.5 ขั้นตอนการศึกษาและดำเนินงานวิจัย

วิทยานิพนธ์นี้มีขั้นตอนการศึกษาและการดำเนินงานวิจัย ดังนี้

1) ศึกษางานวิจัยที่เกี่ยวข้องกับการคัดแยกระดับของเนื้อความและกฎที่ใช้ในการคัดแยกระดับเนื้อความภาษาอังกฤษ

2) ศึกษากระบวนการทำงานของเครือข่ายประสาทเทียม

3) ตั้งสมมติฐานโดยคาดว่าระบบที่ได้จากสมการและถูกคัดแยกด้วยเครือข่ายประสาทเทียม สามารถวิเคราะห์แยกเนื้อความภาษาอังกฤษสำหรับอ่านให้กับนักเรียนไทยได้อย่างเหมาะสมตามความสามารถด้านการอ่านในแต่ละระดับชั้นเรียน และเป็นระบบที่ช่วยผู้ใช้ในการประเมินความยากของเนื้อความสำหรับอ่าน

4) รวบรวมและจัดเก็บคลังข้อมูลเนื้อความและฐานข้อมูลรายการคำศัพท์ เพื่อเป็นข้อมูลอ้างอิงให้ระบบเรียนรู้ว่าเนื้อความและคำศัพท์นี้เป็นเนื้อความและคำศัพท์ของระดับชั้นใด ได้มีการออกแบบระบบให้แบ่งกลุ่มระดับชั้นการเรียนการสอน โดยกำหนดไว้ 4 กลุ่มระดับ

5) ออกแบบโครงสร้างและคุณลักษณะของระบบคัดแยกระดับเนื้อความ และอัลกอริทึม ที่ใช้ในการคัดแยก โดยจะนำไปคำนวณตามแง่มุมทางภาษา (Linguistic feature) ซึ่งจะเข้ามาช่วยในการหาจุดบ่งชี้ความยากสำหรับแต่ละระดับชั้นเรียน เพื่อให้เหมาะสมกับผู้อ่านที่เป็นนักเรียนไทย

6) ใช้เครือข่ายประสาทเทียมเรียนรู้ข้อมูลที่ผ่านการคำนวณจากโครงสร้างและคุณลักษณะที่กำหนดไว้ รวมทั้งปรับค่าและโครงสร้างภายในเครือข่ายประสาทเทียมให้ได้ค่าความผิดพลาดน้อยที่สุด เพื่อระบบจะสามารถระบุความยากของเนื้อความได้อย่างถูกต้องและแม่นยำมากที่สุด

7) ทดลองเพื่อหาความถูกต้องในการทำนายการแบ่งแยกระดับเนื้อความสำหรับอ่านโดยใช้วิธี 5-โฟลด์ คลอสวาลิเดชั่น รวบรวมและสรุปผล

8) เขียนวิทยานิพนธ์

1.6 ประโยชน์ที่คาดว่าจะได้รับ

วิทยานิพนธ์นี้มีประโยชน์ที่คาดว่าจะได้รับ ดังนี้

1) ทำให้ได้อัลกอริทึมที่เหมาะสมสำหรับวัดหรือคัดแบ่งระดับความยากของเนื้อความภาษาอังกฤษสำหรับอ่าน ของนักเรียนไทย

2) สามารถนำโมเดลทางสถิติที่ประยุกต์เข้ากับโครงข่ายประสาทเทียม (Neural Network) มาเป็นพื้นฐานในการวัดระดับความยากของเนื้อความภาษาอังกฤษของนักเรียนไทยในแต่ละระดับชั้นเรียนได้

3) สามารถนำไปประยุกต์ใช้ร่วมกับการเรียนรู้ด้วยสื่ออิเล็กทรอนิกส์สำหรับการแบ่งแยกระดับความยากของเอกสารหรือเนื้อความภาษาอังกฤษในการเรียนรู้ได้อย่างอัตโนมัติ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงคุณลักษณะทั่วไปของงานที่เกี่ยวข้องกับระบบตรวจวัดความยาก-ง่ายของการอ่านสำหรับเนื้อความ รวมทั้งกล่าวถึงทฤษฎีการประยุกต์ข้อมูลทางสถิติที่นำมาใช้พัฒนาระบบตรวจวัดความยาก-ง่ายของการอ่าน

2.1 ระบบตรวจวัดความยาก-ง่ายของเนื้อความ

ในการเรียนการสอนภาษาอังกฤษนั้น เนื้อความสำหรับอ่าน (Reading passage) แต่ละเนื้อความมีระดับความยาก-ง่ายในการอ่านแตกต่างกันไป โดยระดับดังกล่าวจะขึ้นอยู่กับการใช้คำศัพท์ จำนวนพยางค์ของคำศัพท์ และโครงสร้างทางไวยากรณ์ (Grammatical structure) ภายในเนื้อความนั้นๆ ซึ่งโดยทั่วไปเนื้อความสำหรับอ่านเหล่านี้ จะถูกคัดเลือกและกำหนดมาตรฐานโดยผู้เชี่ยวชาญภาษาอังกฤษหรือผู้สอน

ระบบตรวจวัดความยาก-ง่ายของเนื้อความคือ ระบบที่สามารถตรวจสอบและวัดระดับความยากของเนื้อความตามมาตรฐานของระดับความยาก-ง่ายที่ถูกกำหนดไว้โดยคำนวณระดับความยากจากการใช้คำศัพท์ จำนวนพยางค์ของคำศัพท์ และโครงสร้างทางไวยากรณ์ (Grammatical structure) ของเนื้อความสำหรับอ่านแต่ละชิ้น โดยใช้สมการคำนวณระดับความยากที่กำหนดไว้และไม่ใช้การตัดสินใจของบุคคล ซึ่งมีโอกาสที่จะมีอคติ (Bias) หรือความไม่แน่นอน (Inconsistency) และก่อให้เกิดการตัดสินระดับความยากที่ไม่สอดคล้องกับมาตรฐานหรือแตกต่างกันตามมุมมองของบุคคล นอกจากนั้นระบบนี้ยังช่วยลดเวลาในการตรวจสอบระดับของเนื้อความได้อย่างประสิทธิภาพเพราะการตรวจวัดระดับความยาก-ง่ายของเนื้อความสำหรับอ่านนั้นต้องใช้เวลานานจากการที่ต้องอ่านเนื้อความทั้งหมดซึ่งจะเป็นภาระให้แก่ผู้ตรวจวัดโดยทั่วไป ผลลัพธ์ของระบบนี้จะเป็นระดับความยากของเนื้อความ

ตั้งแต่ปีพ.ศ. 2503 จนถึงปัจจุบัน ระบบตรวจวัดความยาก-ง่ายของเนื้อความได้ถูกพัฒนามาแล้วหลายระบบ และมีข้อดีข้อเสียแตกต่างกันไป โดยทั้งหมดใช้สมการคำนวณระดับความยากของตัวบุคคล และให้ผลลัพธ์แตกต่างกันตามแต่มาตรฐานในการวัดระดับความยากของเนื้อความที่ใช้ ระบบส่วนมากจะเน้นที่การพัฒนาสมการในการคำนวณระดับความยากเป็นหลัก เพื่อให้ผลลัพธ์มีความถูกต้องตามแนวความคิดของตัวบุคคลมากที่สุด

2.1.1 Flesch–Kincaid readability test

ระบบตรวจวัดระดับความยากของเนื้อความ Flesch–Kincaid readability test มี 2 สองระดับการตรวจวัดย่อยคือ Flesch Reading Ease และ Flesch–Kincaid Grade Level โดยทั้งสองระบบนี้ใช้สมการในการคำนวณหลักเดียวกันคือ การคำนวณความยาวของคำ และประโยคภายในเนื้อความสำหรับอ่าน แต่แตกต่างกันในการให้น้ำหนักปัจจัย (Weight factor) สำหรับการคำนวณ

2.1.1.1 Flesch Reading Ease ระบบ Flesch Reading Ease เป็นระบบตรวจวัดระดับความยากของเนื้อความที่ได้รับการเผยแพร่มาตั้งแต่ปีพ.ศ. 2491 สำหรับ Flesch Reading Ease คะแนนที่สูงจะชี้ถึงเนื้อความที่ง่ายในการอ่าน และคะแนนที่ต่ำลงจะมีความยากขึ้นตามลำดับ โดยใช้สมการในการคำนวณค่าคะแนนความยาก-ง่ายดังแสดงในสมการ (2.1)

$$\text{Grade} = 206.835 - 1.015\left(\frac{\text{Total words}}{\text{Total sentence}}\right) - 84.6\left(\frac{\text{Total syllables}}{\text{Total words}}\right) \qquad (2.1)$$

โดยกำหนดให้

Total words คือ จำนวนคำศัพท์ทั้งหมด

Total sentence คือจำนวนประโยคทั้งหมด

Total syllables คือ จำนวนพยางค์ทั้งหมด

ผลคะแนนที่ได้ของแต่ละเนื้อความจะถูกแบ่งตามลำดับตามตารางที่ 2.1 โดยกลุ่มคะแนนที่ไม่อยู่ในช่วงที่กำหนดไว้ให้นับเป็นช่วงที่เกี่ยวเนื่องกันในระหว่างระดับ

ตารางที่ 2.1 ลำดับคะแนนในการแบ่งความยากของเนื้อความ

ช่วงคะแนน	ระดับความยาก
90.0–100.0	ระดับสำหรับนักเรียนทั่วไปในช่วงอายุ 11 ปีจะอ่านเข้าใจได้
60.0–70.0	ระดับสำหรับนักเรียนทั่วไปในช่วงอายุ 13-15 ปีจะอ่านเข้าใจได้
0.0–30.0	ระดับสำหรับศึกษามหาวิทยาลัย

จากการทดลอง (A new readability yardstick)[1] ทำการวิจัยโดย Flesch R. Ease พบว่าระดับความยากของนิตยสาร Reader's Digest อยู่ในช่วงคะแนนประมาณ 65 แต้ม และนิตยสาร Time อยู่ในช่วงคะแนนประมาณ 52 แต้มโดยคะแนนสูงที่สุดที่ระบบเคยให้คือประมาณ 120 แต้ม ซึ่งเป็นเนื้อความที่มีประโยคที่มีแต่ คำ 1-2 พยางค์ อยู่จำนวน 2-3 คำ ดังนั้นระบบนี้จะเกิดผลกระทบมากที่สุดจากการปรากฏของคำที่มีจำนวนพยางค์มาก

ระบบนี้ได้ถูกนำไปประยุกต์ใช้กับโปรแกรมอื่นๆ เป็นจำนวนมากเช่น โปรแกรม KWord โปรแกรม IBM Lotus Symphony โปรแกรม Microsoft Office Word โปรแกรม WordPerfect และโปรแกรม WordPro

2.1.1.2 Flesch–Kincaid Grade Level ระบบ Flesch–Kincaid Grade Level เป็นระบบที่ถูกออกแบบมาให้ตรวจวัดระดับความยากที่สอดคล้องกับระดับการเรียนการสอนในโรงเรียนเป็นหลักตั้งแต่ปีพ.ศ. 2518 โดยจะแบ่งระดับของคะแนนเป็น 12 กลุ่ม ตามระดับการเรียนการสอนเกรด 1 ถึงเกรด 12 โดยใช้สมการในการคำนวณค่าคะแนนความยาก-ง่ายดังแสดงในสมการ (2.2)

$$\text{Grade} = 0.39\left(\frac{\text{Total words}}{\text{Total sentence}}\right) + 11.8\left(\frac{\text{Total syllables}}{\text{Total words}}\right) - 15.59 \qquad (2.2)$$

โดยกำหนดให้

Total words คือ จำนวนคำศัพท์ทั้งหมด

Total sentence คือจำนวนประโยคทั้งหมด

Total syllables คือ จำนวนพยางค์ทั้งหมด

ผลการคำนวณของสมการที่ (2.2) อยู่ในค่าซึ่งจะอิงกับระดับชั้นหรือจำนวนปีที่ได้เรียนรู้ภาษาอังกฤษมา เช่น หากได้คะแนน 10.2 จะหมายความว่าเนื้อความนี้จะเหมาะสมสำหรับนักเรียนในระดับเกรด 10 (มัธยม 4) หรือ เรียนภาษาอังกฤษมาเป็นเวลา 10 ปี อ้างอิงจากงานวิจัยของ Kincaid [2][3][4][5] อย่างไรก็ตาม จากการทดลองพบว่าการให้คะแนนอาจให้ผลเป็นค่าติดลบเพราะมีค่าเฉลี่ยความยาวคำและประโยคต่ำมาก เช่น เนื้อความที่มีค่าเฉลี่ยความยาว 5.7 คำต่อประโยคและ1.02 พยางค์ต่อคำได้รับระดับคะแนนอยู่ที่ 1.3 [6]

2.1.2 Gunning fog index

ระบบ Gunning fog index [6][7]เป็นระบบตรวจวัดระดับความยากของเนื้อความอีกผลงานหนึ่ง ระบบนี้ได้รับการพัฒนาตั้งแต่ปี พ.ศ. 2495 ระบบนี้เป็นระบบแรกที่ได้นำความถี่ของการปรากฏคำหลายพยางค์มาใช้ในการคำนวณระดับความยากของเนื้อความ โดยใช้สมการ (2.3) ซึ่งผลคะแนนที่ได้จะอิงจากจำนวนปีที่ได้เรียนรู้ภาษาอังกฤษมา

$$\text{Grade} = 0.4\left(\left(\frac{\text{words}}{\text{sentence}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right) \qquad (2.3)$$

โดยกำหนดให้

words คือจำนวนคำทั้งหมดในเนื้อความ

sentence คือจำนวนประโยคทั้งหมดในเนื้อความ

complex words คือจำนวนคำที่มีความยาวเกิน 3 พยางค์ขึ้นไป

ข้อจำกัดของระบบ Gunning fog index คือบางครั้งคำภาษาอังกฤษที่มีความยาวตั้งแต่ 3 พยางค์ขึ้นไป เช่น elephant (ช้าง /คำนาม) ก็ไม่ใช่คำที่ยากเมื่อเปรียบเทียบกับ mourn (เศร้าโศก / คำกริยา) ดังนั้นหากในเนื้อความ มีคำที่ยากแต่จำนวนพยางค์น้อย จะมีผลทำให้ค่าคะแนนที่ออกมาผิดพลาดและได้คะแนนต่ำกว่าความเป็นจริง

2.1.3 SMOG

ระบบ SMOG grading (Simple Measure of Gobbledygook) [9][10] เป็นระบบตรวจวัดระดับความยากของเนื้อความที่จะให้ผลลัพธ์เป็นค่าคะแนนอิงกับระดับชั้นหรือจำนวนปีที่ได้เรียนรู้ภาษาอังกฤษ โดยได้เปิดให้ใช้งานตั้งแต่ปีพ.ศ. 2512 ระบบ SMOG grading ใช้สมการในการคำนวณที่เน้นไปในการดูจำนวนคำที่มีความยาวตั้งแต่ 3 พยางค์ขึ้นไปจาก 30 ประโยคที่สุ่มขึ้นมาโดยแบ่งเป็น 3 ส่วน คือ ส่วน ละ10 ประโยคที่ต่อเนื่องกัน จากต้นเนื้อความ กลางเนื้อความ และท้ายเนื้อความตามลำดับ และนำไปเปรียบเทียบกับจำนวนประโยคทั้งหมด โดยใช้สมการ (2.4) ในการคำนวณ โดยหากภายในเนื้อความมีจำนวนต่ำกว่า 30 ประโยคจะนำประโยคทั้งหมดมาใช้แทนการสุ่ม

$$\text{Grade} = 1.043\sqrt{30 * \frac{\text{number of polysyllables}}{\text{number of sentences}}} + 3.1291 \qquad (2.4)$$

โดยกำหนดให้

number of polysyllables คือจำนวนคำที่มีความยาวตั้งแต่ 3 พยางค์ขึ้นไป

number of sentences คือจำนวนประโยคทั้งหมดในเนื้อความ

จากการทดลอง พบว่า ระบบ SMOG grading ให้ค่าความถูกต้องที่สูงและมีค่าความคลาดเคลื่อน (Standard error) ของการให้ค่าคะแนนระดับชั้นเพียง 1.5159 เท่านั้น

2.1.4 สรุประบบที่มีอยู่เดิม

ระบบที่กล่าวมาทั้งหมดนั้น เป็นระบบตรวจวัดระดับความยากของเนื้อความ ที่ให้ค่าคะแนนอิงกับความถี่ของจำนวนคำ จำนวนประโยค ความยาวของประโยค และความยาวของพยางค์เป็นหลัก ซึ่งปัจจัยเหล่านี้เป็นปัจจัยที่ไม่ครอบคลุมและสอดคล้องกับระดับความยากที่แท้จริงของเนื้อความ เช่น ประโยคที่มีจำนวนคำมาก ไม่จำเป็นที่จะต้องเป็นประโยคที่ยากเสมอไป เช่นเดียวกันกับ คำที่ยาวที่บางครั้งก็ไม่ใช่คำที่ยากกว่าคำสั้นบางคำ

นอกจากนั้นค่าปัจจัยที่ใช้ในการคำนวณสมการทั้งหมดถูกกำหนดมาให้อ้างอิงกับความสามารถในการอ่านของนักเรียนที่เป็นเจ้าของภาษาอังกฤษทั้งสิ้น หากจะนำระบบใดมาใช้กับนักเรียนชาติอื่นๆ ที่ไม่ใช่เจ้าของภาษาอังกฤษ ค่าปัจจัยที่ใช้ในการคำนวณสมการต้องถูกปรับเปลี่ยนตามมาตรฐานการเรียนการสอนภาษาอังกฤษของประเทศนั้นๆ กล่าวคือ ค่าปัจจัยที่ใช้ในการคำนวณสมการที่ใช้กับประเทศไทยก็จะไม่ให้ความถูกต้องเมื่อนำค่าเดียวกันไปประยุกต์ใช้กับประเทศอื่นๆ ในภูมิภาคเอเชีย ยิ่งไปกว่านั้นค่าปัจจัยเหล่านี้ส่วนมากถูกกำหนดโดยผู้เชี่ยวชาญหรืออาจารย์ผู้สอนภาษาอังกฤษ เช่น ค่า 1.043 และค่า 3.1291 ก็ไม่มีที่มาของการคำนวณค่าเหล่านี้ที่เป็นมาตรฐาน ทำให้การกำหนดค่าเหล่านี้เป็นไปได้ยาก

ดังนั้น หากประยุกต์ใช้การเรียนรู้ด้วยเครื่อง (Machine learning) ระบบจะสามารถกำหนดค่าปัจจัยได้อย่างอัตโนมัติ รวมทั้งหากใช้ฐานข้อมูลหรือคลังข้อมูล (Corpus) ที่รวบรวมเนื้อความสำหรับอ่านภาษาอังกฤษที่ใช้ในการเรียนการสอนของประเทศนั้นๆ ไว้ ระบบก็จะสามารถตรวจวัดระดับความยากได้สอดคล้องกับมาตรฐานระดับความยากของประเทศดังกล่าวได้โดยปริยาย

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง (Machine Learning) คือ การทำให้เครื่องเรียนรู้ได้จากข้อมูลตัวอย่างหรือจากสภาพแวดล้อม จุดมุ่งหมายคือการพัฒนาหรือปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้น เมื่อเรียนรู้แล้วความรู้ที่เรียนได้ จะเก็บไว้ในฐานความรู้ด้วยรูปแบบการแทนความรู้อย่างใดอย่างหนึ่ง เช่น กฎ ฟังก์ชัน เป็นต้น

อัลกอริทึม(Algorithm) เรียนรู้ของเครื่องจะต่างจากอัลกอริทึมที่ถูกโปรแกรมสั่งงานเป็นลำดับขั้นตอนโดยมนุษย์ อัลกอริทึมเรียนรู้ของเครื่องสามารถเรียนรู้ได้เองโดยอัตโนมัติจากตัวอย่างที่ป้อนให้ ดังนั้น อัลกอริทึมเรียนรู้จึงเหมาะกับงานที่ไม่สามารถสั่งงานเป็นขั้นตอนได้ เช่น การรู้จำใบหน้า (Face recognition) ถ้ามีข้อมูลตัวอย่างเป็นจำนวนมากเพียงพอสำหรับใช้เป็นข้อมูลสอน

สามารถใช้ประโยชน์จากข้อมูลเหล่านั้น เพื่อสร้างโมเดล (Model) สำหรับให้เครื่องตัดสินใจหรือแก้ปัญหาได้เองโดยอัตโนมัติ

ข้อมูลตัวอย่างที่จะให้อัลกอริทึมเรียนรู้ แบ่งเป็น 2 ประเภท คือ

1) ตัวอย่างที่มีฉลาก (Labeled example) คือ ตัวอย่างที่มีคำตอบบอกไว้ ว่าเป็นตัวอย่างที่อยู่ในประเภทไหน หรือเป็นตัวอย่างกรณีที่ถูกต้องหรือไม่ถูกต้อง เรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised learning)

2) ตัวอย่างที่ไม่มีฉลาก (Unlabeled example) คือ ตัวอย่างที่ไม่มีคำตอบบอกไว้ว่าเป็นตัวอย่างที่อยู่ในประเภทไหน หรือเป็นตัวอย่างกรณีที่ถูกต้องหรือไม่ถูกต้อง แต่จะถูกกำหนดโดยสภาพแวดล้อม ทั้งนี้ตัวอย่างที่ไม่มีฉลาก อัลกอริทึมจะต้องลองผิดลองถูกเพื่อเรียนรู้เอาเอง เรียกว่าการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) ซึ่งการเรียนรู้จะทำได้ยากกว่า

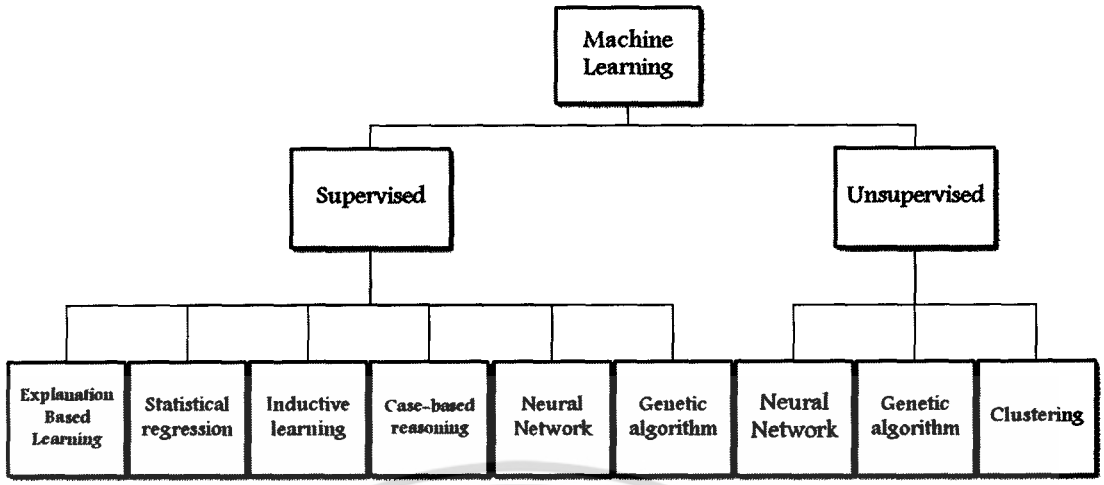
2.2.1 วิธีการเรียนรู้ของเครื่อง

วิธีการเรียนรู้ของเครื่องมีหลายแบบด้วยกัน โดยทั่วไปจะแยกตามข้อมูลตัวอย่างที่จะให้อัลกอริทึมเรียนรู้สามารถแบ่งได้ดังนี้

2.2.1.1 การเรียนรู้แบบมีผู้สอน (Supervised learning) เป็นการเรียนแบบที่มีการตรวจคำตอบเพื่อให้วงจรเครือข่ายปรับตัว ชุดข้อมูลที่ใช้สอนวงจรเครือข่ายจะมีคำตอบไว้คอยตรวจดูว่าวงจรเครือข่ายให้คำตอบที่ถูกหรือไม่ ถ้าตอบไม่ถูก วงจรเครือข่ายก็จะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น (เปรียบเทียบกับคน เหมือนกับการสอนนักเรียนโดยมีครูผู้สอนคอยแนะนำ)

2.2.1.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) เป็นการเรียนแบบไม่มีผู้แนะนำ ไม่มีการตรวจคำตอบว่าถูกหรือผิด วงจรเครือข่ายจะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูลหรือผลลัพธ์ที่ได้ วงจรเครือข่ายจะสามารถจัดหมวดหมู่ของข้อมูลได้ (เปรียบเทียบกับคน เช่น การที่เราสามารถแยกแยะพันธุ์พืช พันธุ์สัตว์ตามลักษณะรูปร่างของมันได้เองโดยไม่มีใครสอน)

ด้วยการเรียนรู้แต่ละประเภทสามารถมีเทคนิคในการเรียนรู้ได้มากกว่าหนึ่ง ซึ่งจะพิจารณาตามความเหมาะสมของชุดข้อมูลที่จะนำมาใช้ เราสามารถแบ่งแยกเทคนิคต่างๆพอสังเขปได้ดังรูปที่ 2.1



รูปที่ 2.1 วิธีการเรียนรู้ของเครื่องในรูปแบบต่างๆ

2.3 โครงข่ายประสาทเทียม (Artificial Neural Network)

2.3.1 ประวัติและความเป็นมาของโครงข่ายประสาทเทียม

Anderson และ Rosenfeld กล่าวว่า McCulloch และ Pitts มีแนวคิดจุดเริ่มต้นที่ว่า ถ้าอินพุตหนึ่งไปยังนิวรอนหนึ่งที่มีค่าใหญ่กว่าค่าเริ่มต้น หน่วยยิง (Unit fire) จะทำการยิง ซึ่งเป็นแนวคิดที่ใช้กันในปัจจุบัน และถูกนำมาใช้กว้างขวางมากที่สุดในวงจรตรรกะ ซึ่ง Garson ได้ลำดับพัฒนาการของโครงข่ายประสาทเทียม ต่อมา Hebb (1949) ได้นำเสนอการค้นหาโมเดลทางชีววิทยาเพื่ออธิบายความเป็นไปได้ว่าในโครงสร้างของเพอร์เซปตรอนเกิดการเรียนรู้ได้อย่างไร ซึ่งเป็นส่วนสำคัญในการพัฒนาทฤษฎีของเครือข่ายประสาทเทียม และต่อมารู้จักกันดีในนาม "Hebbian learning rule" Rosenblatt [12] ได้ตีพิมพ์ผลงานประดิษฐ์เพอร์เซปตรอนขึ้นใช้ในงานวิจัยเกี่ยวกับการเรียนรู้ และได้คิดค้นกฎการเรียนรู้ให้กับโครงข่ายประสาท แต่ว่ายังแก้ปัญหาได้เฉพาะปัญหาที่ไม่ซับซ้อน กล่าวคือข่ายงานที่เขาสร้างขึ้นนั้นไม่สามารถแยกแยะรูปแบบของข้อมูลใหม่ที่ไม่เคยเห็นมาก่อน Widrow ได้พัฒนาแนวคิดของ McCulloch และ Pitts ไปเป็นโครงข่ายชั้นเดียวที่เรียกว่า adaptive linear neuron (Adaline) และต่อมาได้ปรับระบบการเรียนให้มีความซับซ้อนมากกว่าเพอร์เซปตรอนของ Rosenblatt โดยมีโครงข่ายเป็นสองชั้นเรียกว่าเป็น multiple adaptive linear neuron (Medaline) ต่อมา Widrow และ Hoff ได้พัฒนากฎการเรียนรู้ขึ้นมาใหม่ ซึ่งยังใช้กันอยู่ในปัจจุบันโดยเรียกกฎนี้ว่า "Widrow-Hoff learning rule" แต่ยังคงใช้ได้กับโครงข่ายอย่างง่าย จากนั้น Minsky และ Papert ได้เสนอโครงข่ายแบบใหม่เพื่อให้แก้ปัญหาได้ซับซ้อนมากขึ้น แต่ยังไม่ประสบความสำเร็จในการพัฒนากฎการเรียนรู้ให้ดีขึ้น ต่อมา Kohonen ได้คิดค้นโครงข่ายแบบใหม่ที่เรียกว่า Kohonen networks จนเมื่อ ค.ศ. 1980 เป็นยุคที่คอมพิวเตอร์ได้รับการพัฒนาให้ทำงานได้เร็วขึ้น ซึ่งผลให้มี

การคิดค้นกระบวนการวิธีเรียนรู้ใหม่ๆ เกิดขึ้น ที่สำคัญคือกระบวนการวิธีแพร่กระจายแบบย้อนกลับ (Back propagation algorithm) โดยนำมาใช้กับโครงข่ายประสาทแบบหลายชั้น

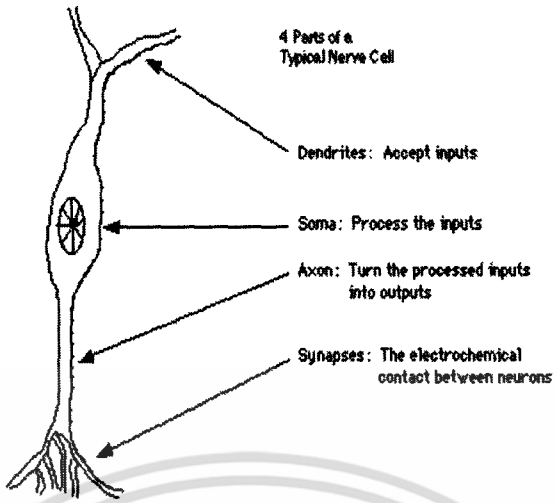
2.3.2 โครงสร้างและการทำงานของโครงข่ายประสาทของมนุษย์

เนื่องจากโครงข่ายประสาทเทียมนั้นเป็นการประยุกต์แนวคิดและหลักการทำงานของโครงข่ายประสาทของมนุษย์ การที่จะศึกษาและเข้าใจการทำงานจนถึงขั้นของการประยุกต์ใช้นั้น จึงควรทำความเข้าใจเกี่ยวกับโครงข่ายประสาทมนุษย์เสียก่อน โครงข่ายประสาทนั้นประกอบขึ้นด้วยส่วนสำคัญ 3 ส่วน คือ

2.3.2.1 เซลล์ประสาท (Neuron or soma) ในสมองของมนุษย์นั้นประกอบไปด้วยเซลล์ประสาทหรือเรียกอีกอย่างว่าหน่วยประมวลผล (Computing or processing elements or elementary nerve cell) ซึ่งมีจำนวนมาก คือประมาณ 1011 ยูนิต มีการเชื่อมโยงกันอย่างหนาแน่นเป็นโครงข่าย มีความหนาแน่นของการเชื่อมโยง 104 ต่อหนึ่งนิวรอน เซลล์ประสาทมีส่วนประกอบเหมือนเซลล์ทั่วไป เช่น ไมโตคอนเดรีย นิวเคลียสไซโตพลาสซึม เป็นต้น

2.3.2.2 ใยประสาท (Nerve fiber) ประกอบไปด้วยสำคัญ 2 ส่วน คือ เดนไดรท์ (Dendrites) ทำหน้าที่รับความรู้สึกจากเส้นใยประสาท โดยส่งข้อมูลเป็นสัญญาณไฟฟ้า (Electrical signal) แต่การทำงานของนิวรอนนั้นก็อาจเกิดขึ้นเนื่องจากกระบวนการทางชีวเคมี (Chemical signal) เข้าสู่เซลล์ประสาทได้ด้วยเดนไดรท์ มีแขนงสั้น ๆ มากกว่าหนึ่งแขนง ไม่มีเยื่อไมอีลินหุ้มเส้นผ่านศูนย์กลางไม่เท่ากันโดยตลอด ส่วนที่สองเรียกว่าแอกซอน (Axon) ทำหน้าที่ส่งกระแสประสาทออกจากตัวเซลล์ มักเป็นแขนงยาวเพียง 1 แขนงเส้นผ่านศูนย์กลางเท่ากัน โดยตลอดและมีเยื่อไมอีลินหุ้ม เชื่อมนิวรอนเข้าด้วยกันเป็นระบบโครงข่าย

2.3.2.3 ไซแนปส์ (Synapse) เป็นส่วนที่อยู่ระหว่างเอกซอนของนิวรอนหนึ่งกับเดนไดรท์ของนิวรอนที่อยู่ข้างเคียง จุดที่สัมผัส (Impinge) กันของเอกซอนกับเดนไดรท์มีอยู่มากถึง 1014 แห่งด้วยกันไซแนปส์นี้เป็นจุดของการถ่ายทอดกระแสความรู้สึกโดยอาจอยู่ในรูปของสารเคมี (Chemical synapse) หรือถ่ายทอดในรูปของกระแสไฟฟ้า (Electrical synapse) นอกจากนั้น ยังมีส่วนประกอบย่อยคือเซลล์ชวาน (Schwann cell) ซึ่งเป็นเซลล์บางๆ ติดต่อกับเยื่อไมอีลินและทำหน้าที่สร้างเยื่อไมอีลินที่หุ้มแอกซอนนั่นเอง เยื่อไมอีลินนี้เป็นสารพวกฟอสโฟไลปิด (Phospholipids) มีคุณสมบัติเป็นฉนวนไฟฟ้าทำให้กระแสประสาทเคลื่อนที่ได้รวดเร็วขึ้นด้วยการกระโดดข้ามไปยังคอคอดของแอกซอนที่ไม่มีเยื่อไมอีลินหุ้ม ซึ่งคอคอดนี้เรียกว่าโหนดออฟแรนเวียร์ (Node of ranvier) รายละเอียดดังที่กล่าวมาแสดงดังรูปที่ 2.2



รูปที่ 2.2 โครงข่ายประสาทเทียม

ด้วยลักษณะโครงสร้างและการทำงานโครงข่ายประสาทมนุษย์ดังกล่าว จึงทำให้สมองของมนุษย์สามารถทำงานหลายๆ งานที่ซับซ้อนพร้อม ๆ กันได้เร็วกว่าคอมพิวเตอร์ ทั้งยังสามารถจดจำและเรียนรู้จากประสบการณ์ในอดีตและนำมาปรับใช้กับสถานการณ์ปัจจุบัน

2.3.3 โครงสร้างของโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมถูกพัฒนาขึ้น เพื่อเลียนแบบความสามารถของระบบประสาททางชีวภาพ ซึ่ง Roa อ้างถึงใน [13] ระบุว่าโครงสร้างของข่ายงานระบบประสาทนั้นมีด้วยกันหลายโครงสร้างแต่ทุก ๆ โครงสร้างก็มีข้อดีที่สำคัญ ๆ ร่วมกัน เช่นลักษณะที่สำคัญที่สุดของข่ายงานระบบประสาท คือความสามารถในการประมาณฟังก์ชันต่อเนื่องแบบไม่เป็นเชิงเส้นใดๆ (Nonlinear continuous function) ในระดับความแม่นที่ต้องการได้ ด้วยคุณลักษณ์นี้ข่ายงานระบบประสาท จึงถูกนำไปใช้ในการหาแบบจำลองของระบบไม่เชิงเส้นเพื่อนำไปประยุกต์ใช้ในการสังเคราะห์หาตัวควบคุมต่อไป ข่ายงานระบบประสาทสามารถมีหลายสัญญาณเข้าและหลายสัญญาณออก จึงสามารถนำไปประยุกต์ใช้กับระบบหลายตัวแปร (Multivariable system) ได้ง่าย นอกจากนั้น Herz อ้างถึงใน [14] ยังได้กล่าวถึงข้อดีของการศึกษาเพื่อเลียนแบบระบบประสาทไว้ 5 ประการ คือ

ก) ความแข็งแกร่ง เนื่องจากเซลล์ประสาทในสมองของมนุษย์มีการตายทุกวันซึ่งถือว่าเป็นเรื่องปกติเช่นเดียวกับการตายของเซลล์ผิวหนัง แต่ก็ไม่มีผลกระทบต่อประสิทธิภาพการทำงานของสมอง นั่นคือ ถ้าส่วนใดส่วนหนึ่งของนิวรัลเน็ตเวิร์คเสียไป ก็ยังทำงานต่อได้

ข) ความยืดหยุ่นเนื่องจากมนุษย์มีความสามารถในการเรียนรู้ ทำให้สามารถปรับสภาพให้เข้ากับสิ่งแวดล้อมต่าง ๆ ซึ่งเป็นสิ่งที่ต้องมีการปรับปรุงอยู่เสมอของระบบดิจิตอลคอมพิวเตอร์

ค) ความสามารถในการจัดการบ่งชี้ข้อมูล นิวรัลเน็ตเวิร์คมีความสามารถในการจัดการกับข้อมูลที่คลุมเครือ ไม่สมบูรณ์หรือขัดแย้งกันได้สูง

ง) ความสามารถในการทำงานคู่ขนาน จากชั้นหนึ่งซึ่งประกอบไปด้วยโหนดหลายโหนดไปอีกชั้นหนึ่งซึ่งประกอบไปด้วยโหนดหลายโหนดเช่นเดียวกับการทำงานของระบบประสาทมนุษย์

จ) มีขนาดเล็กและใช้พลังงานน้อย

ในปี 1989 Wassermann ก็ได้กล่าวถึงคุณสมบัติที่เป็นข้อดีของโครงข่ายประสาทเทียมไว้ 2 ประการคือ การที่นิวรัลเน็ตเวิร์คสามารถเรียนรู้จากชุดการสอนที่ถูกป้อนให้เรียนรู้ และอีกประการหนึ่งคือการที่นิวรัลเน็ตเวิร์คสามารถระลึกได้ทั้งชุดการสอนและชุดการทดสอบ หรือชุดทั่วไปได้ดีในระดับที่ยอมรับได้ โดยชุดการทดสอบจะมีความแตกต่างจากชุดที่นิวรัลเน็ตเวิร์คได้เรียนรู้อยู่บ้าง ซึ่งความแตกต่างนี้เรียกว่าสิ่งปนเปื้อน หรือสิ่งบิดเบือนจากข้อมูลประเภทนั้น ๆ หลังจากนั้นในปี 1992 ได้มีการกล่าวถึงประโยชน์จากการประยุกต์นิวรัลเน็ตเวิร์ค 3 ประการ คือ

ก) ความทนทนจากการที่สามารถออกแบบให้มีจำนวนโหนดที่มากมายได้ และหากโหนดใดโหนดหนึ่ง ถูกทำลาย ก็จะไม่ทำให้เน็ตเวิร์คหยุดการทำงานไป

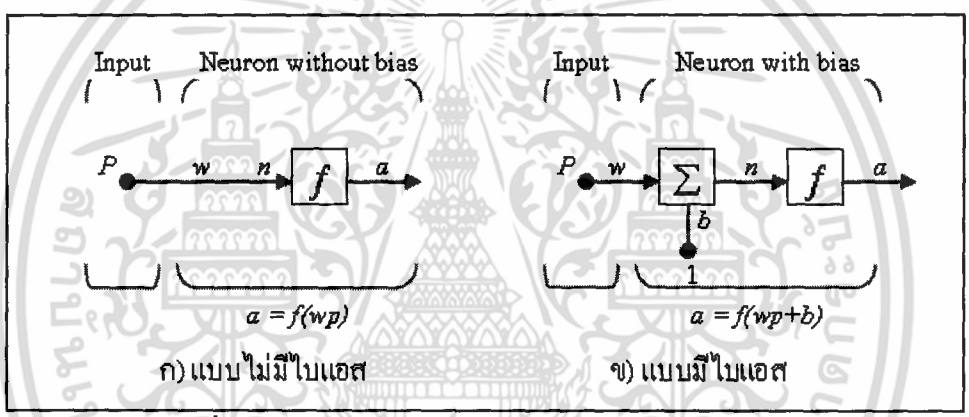
ข) ความสามารถในการปรับตัวให้เข้ากับสิ่งแวดล้อมใหม่ ๆ ได้

ค) ความสามารถในการบ่งชี้ลักษณะทั่วไป ซึ่งเป็นผลมาจากความสามารถในการปรับตัวได้ดีจึงทำให้นิวรัลเน็ตเวิร์คสามารถจัดการกับข้อมูลที่ไม่สมบูรณ์ หรือมีสิ่งปนเปื้อนในอาการประเภทเดียวกันได้เป็นอย่างดี

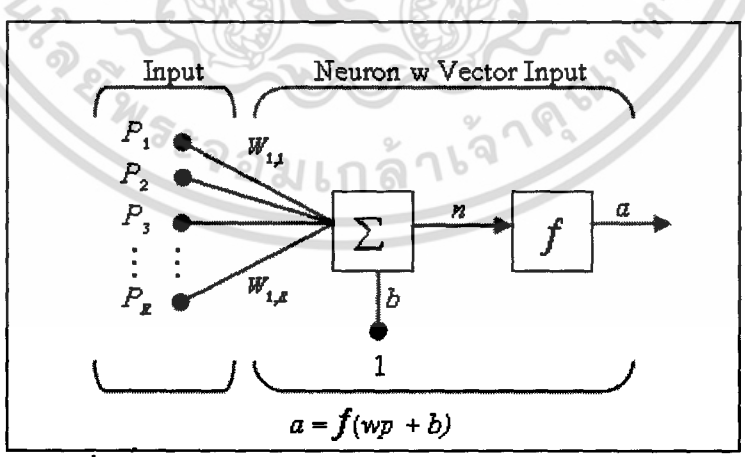
ในการเลียนแบบโครงข่ายประสาทนั้น เริ่มตั้งแต่โครงสร้างของโครงข่ายประสาท รวมทั้งการทำงานซึ่งลักษณะภาพรวมของโครงข่ายประสาทเทียมถูกกำหนดด้วยองค์ประกอบ 3 ส่วน ส่วนที่หนึ่งเป็นรูปแบบการเชื่อมโยงระหว่างนิวรอนหรือเรียกส่วนนี้ว่าสถาปัตยกรรมของโครงข่าย (Network architecture) ส่วนที่สองเป็นการกำหนดการเชื่อมโยงหรือเรียกส่วนนี้ว่า ขั้นตอนวิธีการปรับสอนหรือการเรียนรู้ และส่วนที่สามเป็นการกำหนดข้อมูลส่งออกโดยแอกติเวชันฟังก์ชัน (Activation function) ซึ่ง Jain และ Mao (1994) กล่าวว่าวิธีการกำหนดให้มีการเชื่อมโยงแต่ละจุดของข่ายงานจะทำให้เกิดโครงข่ายประสาทเทียมต่างชนิดกันและจะนำมาซึ่งรูปแบบการเรียนรู้ที่แตกต่างกันโครงข่ายดังกล่าวถูกพัฒนาขึ้นจากแบบจำลองทางคณิตศาสตร์ของการรู้จำในมนุษย์ (Human recognition) หรือนิวรอนทางชีววิทยาโดยตั้งอยู่บนพื้นฐานสี่ประการ ประการแรกคือ การประมวลผลสารสนเทศเกิดขึ้นที่หน่วยพื้นฐานหลาย ๆ หน่วย เรียกว่า นิวรอน (Neuron) ประการที่สองคือ สัญญาณถูกส่งระหว่างนิวรอนผ่านส่วนเชื่อมโยง (Connection link) ประการที่สามคือส่วนเชื่อมโยงมีความสัมพันธ์กับค่าน้ำหนักคูณด้วยค่าสัญญาณที่ส่ง และประการที่สี่แต่ละนิวรอนนำค่าแอกติเวชันฟังก์ชัน ซึ่งโดยปกติเป็นฟังก์ชันที่ไม่ใช่ฟังก์ชันเชิงเส้นไปใช้กับเนตอินพุต ในปี 1998

Garson ได้กล่าวสอดคล้องกับ Lin และ Lee (1996) ว่าการออกแบบโครงข่ายประสาทนั้นครอบคลุม 3 ประการคือ การออกแบบส่วนประกอบในหน่วยคำนวณ การออกแบบการเชื่อมโยงและโครงสร้าง และการออกแบบ กฎการเรียนรู้ โครงสร้างการทำงานของโครงข่ายประสาทเทียมประกอบด้วยชั้นของข้อมูลเบื้องต้น 3 ชั้น คือ ชั้นของข้อมูลป้อนเข้า (Input layer) ชั้นซ่อน (Hidden l layer) และชั้นผลลัพธ์ (Output layer) นักการศึกษาได้พัฒนาโครงข่ายประสาทขึ้นมา และสามารถแบ่งประเภทของโครงข่ายประสาทเทียมตามความซับซ้อน หรือจำนวนชั้นของข้อมูลในชั้นซ่อน ได้เป็น 2 ประเภท คือ

ก) โครงข่ายประสาทเทียมอย่างง่าย (Single layer neural network) เป็นโครงข่ายที่มีชั้นซ่อนเพียงหนึ่งยูนิต ประกอบด้วยเวกเตอร์ของตัวแปรด้านเข้า เมตริกซ์ค่าถ่วงน้ำหนัก เวกเตอร์ตัวแปรด้านนอกและไบแอส (Bias) โครงสร้างของโครงข่ายประสาทเทียมอย่างง่ายแสดงได้ดังรูปที่ 2.5



รูปที่ 2.3 โครงข่ายประสาทเทียมแบบหนึ่งหน่วยแบบง่าย



รูปที่ 2.4โครงข่ายประสาทเทียมแบบหนึ่งแบบหลายอินพุต

จากรูปที่ 2.4 โครงข่ายประสาทเทียบแบบหลายอินพุต มีอินพุต p หลายตัว อินพุตแต่ละตัวถูกคูณด้วยเวกเตอร์น้ำหนัก w แล้วป้อนให้กับฟังก์ชันถ่ายโอน f เป็นเอาท์พุต a ดังสมการ

$$n = w_{1,1}p_1 + w_{1,2}p_2 + \ldots + w_{1,R}p_R + b \qquad (2.5)$$

$$a = f(wp+b)$$

ความสัมพันธ์ระหว่างอินพุตและเอาต์พุตของทรานเฟอร์ฟังก์ชัน เป็นดังนี้

แบบ Linear มีความสัมพันธ์ $a = n$

แบบ Log-Sigmoid มีความสัมพันธ์ $a = \frac{1}{1+e^{-n}}$

แบบ Hard Limit มีความสัมพันธ์ $a = 0 \quad n < 0$

$a = 1 \quad n > 0$

a = purelin(n) — Linear Transfer Function

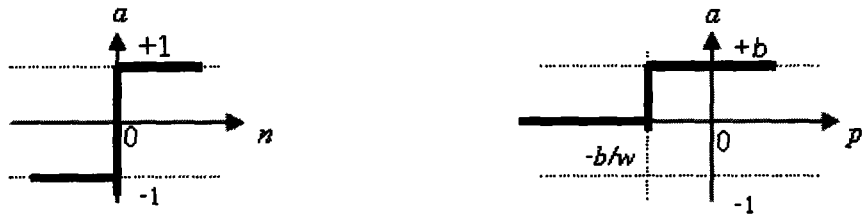
a = purelin(wp+b) — Single-Input *purelin* Neural

รูปที่ 2.5 Linear transfer function

a = logsig(n) — Log-Sigmoid Transfer Function

a = logsig(wp+b) — Single-Input *logsig* Neural

รูปที่ 2.6 Log-Sigmoid transfer function

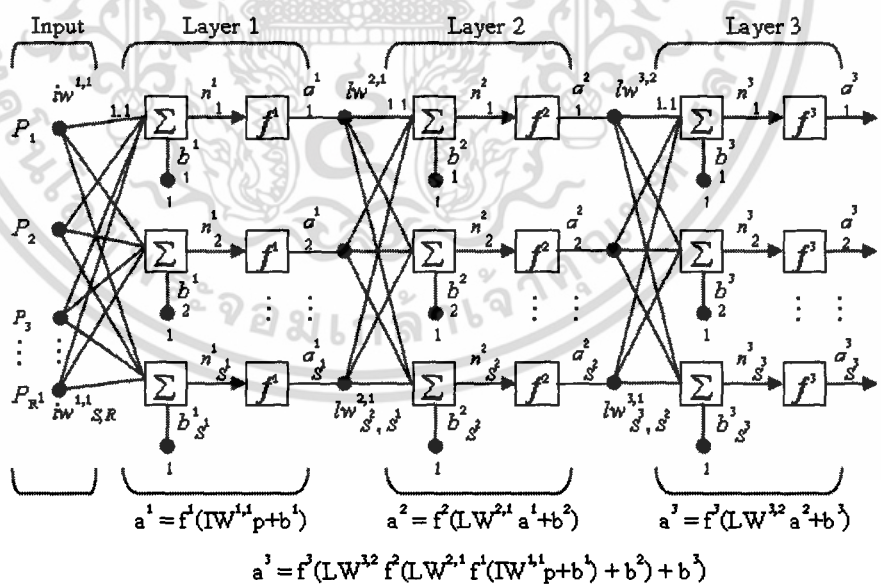


$a = hardlim(n)$ Hard Limit Transfer Function

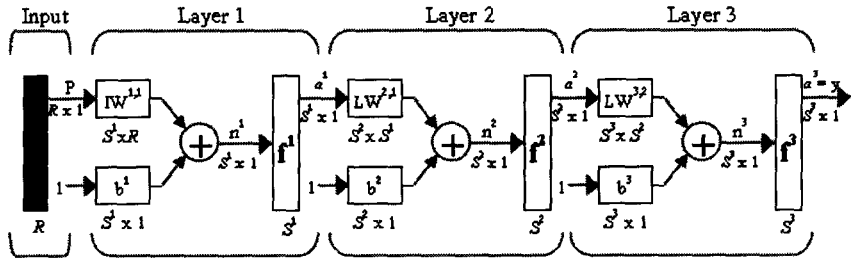
$a = hardlim(wp+b)$ Single-Input *hardlim* Neural

รูปที่ 2.7 Hard Limit transfer function

ข) โครงข่ายประสาทเทียมแบบหลายชั้น (Multi-layer neural network) ประกอบด้วย ส่วนรับข้อมูล (Sensory unit or source node) ซึ่งโดยทั่วไปเรียกว่าชั้นตัวแปรด้านเข้า ส่วนการคำนวณ (Computation nodes) เรียกว่า ชั้นซ่อน (Hidden layer) ซึ่งอาจจะมีหนึ่งชั้นหรือมากกว่าก็ได้ (ในกรณีที่มีหนึ่งชั้น ก็ต้องมีจำนวนมากกว่าหนึ่งยูนิต) และส่วนข้อมูลออกจากส่วนการคำนวณเรียกว่าชั้นตัวแปรด้านนอก (Output layer) รูปที่ 2.8 แสดงโครงสร้างของโครงข่ายประสาทเทียมแบบสามชั้น และรูปที่ 2.9 แสดงโครงสร้างของโครงข่ายประสาทเทียมแบบสามชั้นในรูปเมทริกซ์



รูปที่ 2.8 โครงข่ายประสาทเทียมแบบสามชั้น



รูปที่ 2.9 โครงข่ายประสาทเทียมแบบสามชั้นในรูปเมทริกซ์

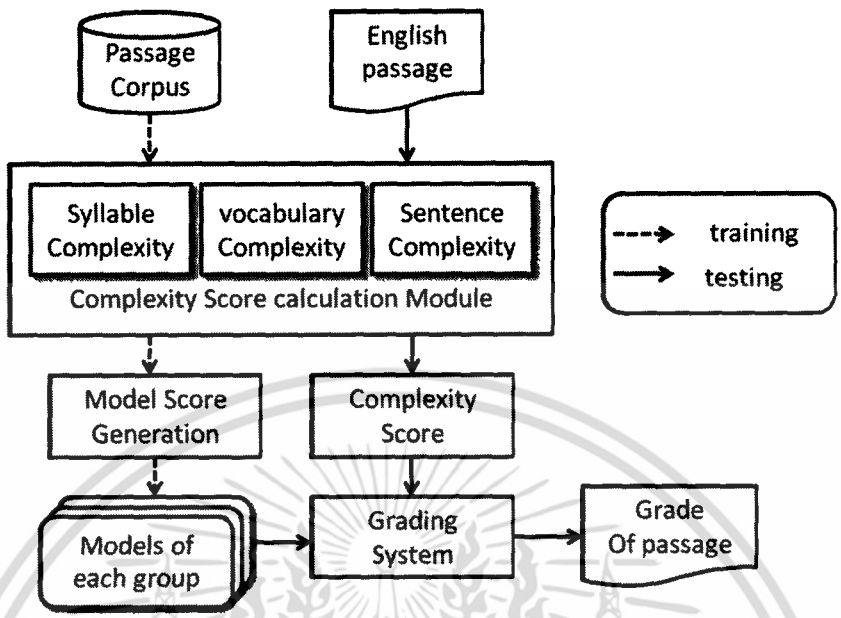
บทที่ 3
ระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านโดยใช้เครือข่ายประสาทเทียม

วิทยานิพนธ์ฉบับนี้เป็นงานวิจัยที่นำเสนออัลกอริทึมสำหรับระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านโดยใช้เครือข่ายประสาทเทียม (Passage Grading System Using Neural Network)

เนื้อหาในบทที่ 3 นี้จะกล่าวถึงองค์ประกอบ การเก็บฐานข้อมูลและกระบวนการทำงานของระบบคัดแบ่งที่ถูกพัฒนาขึ้น รวมทั้งอัลกอริทึมและสมการที่ใช้ในการคำนวณหาระดับของเนื้อความภาษาอังกฤษสำหรับอ่าน

ระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านโดยใช้เครือข่ายประสาทเทียมเป็นระบบที่ช่วยผู้ใช้ในการประเมินความยากของเนื้อความสำหรับอ่าน(Reading passage) ภาษาอังกฤษสำหรับนักเรียน โดยใช้การเรียนรู้ด้วยเครื่อง (Machine learning) จากข้อมูลภายในคลังข้อมูลเนื้อความ (Passage corpus) ผลลัพธ์ของระบบคือความยากของเนื้อความ โดยแบ่งตามกลุ่มระดับชั้นการเรียนการสอน ซึ่งกำหนดไว้เป็น 4 กลุ่มระดับ (Grade group) เรียงจากระดับที่ง่ายที่สุดไประดับยากที่สุดตามลำดับ คือ ประถมต้น (ป.1-3) ประถมปลาย (ป.4-6) มัธยมต้น (ม.1-3) มัธยมปลาย (ม.4-6)

ภายในระบบ การคำนวณความยากของเนื้อความจะใช้แง่มุมทางภาษา (Linguistic feature) เข้ามาช่วยในการหาจุดบ่งชี้ความยาก ซึ่งระบบนี้จะเน้นไปที่ ความซับซ้อนของวากยสัมพันธ์ (Syntactic complexity) หรือโครงสร้างของประโยค (Sentence structure) และความยากของคำศัพท์ที่ใช้ในเนื้อความ (Difficulty of vocabulary) เป็นหลัก และระบบจะประยุกต์การใช้ค่าสถิติด้วยเครือข่ายประสาทเทียม (Neural Network) เพื่อปรับจูนค่าความสมดุลของคะแนนที่ได้ การปรับความสมดุลของคะแนนจะสร้างค่าความน่าจะเป็นของต้นแบบ เพื่อใช้ในการเปรียบเทียบกับเนื้อความอื่นๆ ที่ต้องการทราบระดับความยากของเนื้อความ



รูปที่ 3.1 ภาพรวมกระบวนการทำงานของระบบ

3.1 ฐานข้อมูลสำหรับคัดแบ่งระดับเนื้อความ

3.1.1 คลังข้อมูลเนื้อความ

คลังข้อมูลเนื้อความ (Passage corpus) คือฐานข้อมูลที่บรรจุเนื้อความสำหรับอ่านภาษาอังกฤษที่ใช้ในการเรียนการสอนภาษาอังกฤษภายในชั้นเรียนของโรงเรียน การจัดเก็บเนื้อความดังกล่าวต้องจัดเก็บพร้อมทั้งระดับชั้นที่ใช้ในการเรียนของเนื้อความนั้นๆ เพื่อเป็นจุดอ้างอิงให้ระบบ เรียนรู้ว่าเนื้อความนี้เป็นเนื้อความของระดับชั้นใด โดยสำหรับระบบนี้ ได้แบ่งกลุ่มระดับชั้นการเรียนการสอนซึ่งกำหนดไว้ 4 กลุ่มระดับ (Grade group) เรียงจากระดับที่ง่ายที่สุดไประดับยากที่สุดตามลำดับ คือ ประถมต้น (ป.1-3) ประถมปลาย (ป.4-6) มัธยมต้น (ม.1-3) มัธยมปลาย (ม.4-6)

การรวบรวมเนื้อความสำหรับอ่านนั้น มีข้อแม้สำคัญอยู่คือ ต้องเป็นเนื้อความที่ได้รับการรับรองว่าเหมาะสมกับระดับชั้นนั้นอย่างแท้จริง รวมทั้งต้องเป็นเนื้อความที่ใช้จริงในการเรียนการสอนในชั้นเรียนภาษาอังกฤษภายในโรงเรียนที่ได้รับการยอมรับจากหน่วยงานรัฐบาลที่รับผิดชอบของประเทศนั้นๆ เช่น ในประเทศไทยต้องเป็นโรงเรียนที่ได้รับการยอมรับจากกระทรวงศึกษาธิการ ทั้งนี้การรวบรวมเนื้อความสำหรับอ่านต่อกลุ่มระดับ หากมีปริมาณมาก ก็ย่อมให้ความถูกต้องสูงขึ้นตามไปด้วย

3.1.2 ฐานข้อมูลรายการคำศัพท์

ฐานข้อมูลรายการคำศัพท์ (Vocabulary list) คือรายการคำศัพท์ภาษาอังกฤษที่นักเรียนในแต่ละระดับชั้นควรรู้ กล่าวคือรายการคำศัพท์พื้นฐานสำหรับนักเรียนตามระดับชั้น โดยทั่วไปหน่วยงานรัฐบาลที่รับผิดชอบเรื่องการศึกษาจะเป็นผู้กำหนดรายการคำศัพท์ดังกล่าวและให้ผู้สอนในโรงเรียนได้นำคำศัพท์เหล่านี้ไปใช้ในการเรียนการสอน

ฐานข้อมูลรายการคำศัพท์สำหรับระบบนี้ต้องถูกจัดเก็บไว้คู่กับระดับชั้นเพื่อใช้เป็นดัชนีในการเทียบคำศัพท์ที่พบในเนื้อความเพื่อบ่งบอกถึงระดับความยากของคำศัพท์ภายในเนื้อความ

3.2 กระบวนการคัดแบ่งระดับเนื้อความ

ในกระบวนการการคัดแบ่งระดับเนื้อความนั้น ถูกแบ่งเป็น 2 ส่วนคือ ขั้นตอนการเรียนรู้ (Learning phase) และขั้นตอนการคัดแบ่ง (Grading phase) โดยขั้นตอนการเรียนรู้จะเป็นขั้นตอนที่รับข้อมูลรับเข้า (Input) จากคลังข้อมูลเนื้อความเพื่อสร้างโมเดลระดับความยากของเนื้อความ ส่วนขั้นตอนการคัดแบ่งนั้น จะรับข้อมูลอินพุต (Input) ที่เป็นเนื้อความเดี่ยว (Single passage) เพื่อประเมินว่าเนื้อความนั้นอยู่ในกลุ่มระดับชั้นใด

ทั้งขั้นตอนการเรียนรู้และขั้นตอนการคัดแบ่งนั้น ข้อมูลจะถูกนำไปคำนวณตามแง่มุมทางภาษา (Linguistic feature) ซึ่งจะเข้ามาช่วยในการหาจุดบ่งชี้ความยากสำหรับแต่ละระดับ โดยระบบนี้ได้กำหนดแง่มุมทางภาษาไว้ 3 คุณลักษณะ คือ 1) ความยาวของพยางค์ต่อคำ 2) ความยากของคำศัพท์ และ 3) ความซับซ้อนของประโยค ระบบจะคำนวณแง่มุมทางภาษาแยกกันออกเป็น 3 ค่าคะแนน โดยแต่ละค่าคะแนนจะสะท้อนความยากทางภาษาแตกต่างกันออกไป

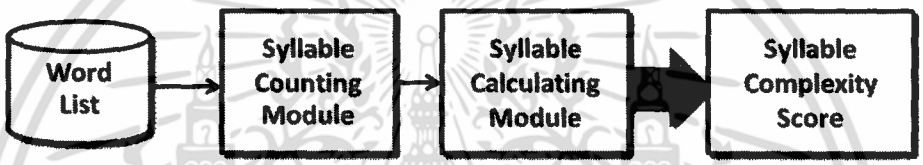
แต่สำหรับขั้นตอนการเรียนรู้ ผลลัพธ์ค่าคะแนน ทั้ง 3 จะถูกนำไปเข้ากระบวนการการคำนวณค่าความน่าจะเป็นด้วยเครือข่ายประสาทเทียมเพื่อปรับค่าคะแนนเหล่านั้นให้กลายเป็นโมเดลระดับความยากของเนื้อความ เพื่อภายหลังจะถูกนำไปเป็นโมเดลในการเทียบกลุ่มระดับชั้นกับเนื้อความจากขั้นตอนการคัดแบ่งด้วยกระบวนการประเมินกลุ่มระดับชั้นต่อไป

ก่อนการคำนวณ ระบบจะต้องแบ่งคำภายในเนื้อความโดยใช้อักขระว่าง (White space) เป็นตัวบ่งขอบเขตของคำ เนื่องจากหน่วยคำศัพท์เป็นหน่วยทางภาษา (Language unit) ซึ่งเป็นพื้นฐานสำหรับการคำนวณค่าในแง่มุมทางภาษาทุกค่า

3.2.1 การคำนวณความยาวของพยางค์ต่อคำ

พยางค์เป็นหน่วยพื้นฐานสำหรับการเรียนรู้ในด้าน ฟัง พูด อ่าน เขียน ในทุกๆภาษา ซึ่งพยางค์จะประกอบไปด้วยหน่วยเสียงที่สามารถรวมกันออกมาเป็นคำศัพท์ได้

ความยาวของพยางค์ต่อคำเป็นแนวคิดที่ได้จากการศึกษางานวิจัยอื่นๆ ที่มีอยู่แล้ว ซึ่งเชื่อว่า คำที่มีจำนวนพยางค์มากกว่าย่อมมีความเป็นไปได้สูง ว่าคำนั้นจะยากกว่าคำที่มีจำนวนพยางค์สั้นกว่า โดยที่จำนวนพยางค์จะสะท้อนความยากของเนื้อความภาษาอังกฤษสำหรับอ่านของนักเรียนแต่ละระดับชั้น อีกทั้งยังพบว่าความยาวของพยางค์ต่อคำเป็นลักษณะทางภาษาแบบหนึ่งที่ทุกระบบที่มีมาก่อนนำมาใช้ กระบวนการการคำนวณความยาวของพยางค์ต่อคำแสดงในรูปที่ 3.2



รูปที่ 3.2 กระบวนการการคำนวณความยาวของพยางค์ต่อคำ

สำหรับการนับพยางค์ของระบบสามารถแบ่งเป็น 4 ขั้นตอนได้ดังนี้

1) นับจำนวนสระในคำศัพท์
2) ลบจำนวนสระที่ไม่ออกเสียง เช่น “e” ที่ปรากฏอยู่ท้ายสุดของคำ
3) เสียงที่มีการออกควบกล้ำจะนับเป็น 1 เสมอ เช่น “th” , “io”
4) จำนวนของเสียงสระจะเป็นจำนวนพยางค์ของคำด้วยเช่นกัน

จำนวนของพยางค์เมื่อเปรียบเทียบกับการออกเสียงก็คือจำนวนของเสียงสระที่ได้ยินจากการออกเสียง ยกตัวอย่างเช่น “cake” มีสระปรากฏทั้งหมด 2 ตัว แต่ “e” ที่ปรากฏอยู่ท้ายคำไม่ออกเสียงจึงไม่มีการนับ ทำให้สระที่ต้องออกเสียงเหลือเพียง 1 จึงทำให้คำศัพท์นี้มีเพียง 1 พยางค์ หรือ “outside” มีสระปรากฏทั้งหมด 4 ตัว แต่ “e” ที่ปรากฏอยู่ท้ายคำไม่ออกเสียงจึงไม่มีการนับและ “ou” เป็นสระควบกล้ำจึงนับจำนวนสระควบกล้ำเป็น 1 ดังนั้นคำศัพท์นี้จึงปรากฏสระที่จำเป็นต้องออกเสียง 2 ตัว คือ “ou” และ “i” ทำให้มีจำนวนพยางค์เท่ากับ 2 พยางค์ นอกจากขั้นตอนในการนับพยางค์เบื้องต้นแล้วยังมีข้อจำกัดประเภทของพยางค์และการแบ่งคำให้เป็นพยางค์เข้ามาช่วยในการนับพยางค์ของคำเพื่อให้เกิดความถูกต้องมากยิ่งขึ้น

ประเภทของพยางค์คือ ลักษณะการออกเสียงและไม่ออกเสียงของพยัญชนะหรือสระที่เป็นไปตามกฎทางภาษามีความแตกต่างกันโดยสามารถแบ่งประเภทของพยางค์ได้ 6 ประเภท ได้แก่

1) พยางค์แบบปิด คือ คำศัพท์ที่ปรากฏสระเพียงหนึ่งเดียวและจบคำศัพท์นั้นด้วยพยัญชนะ เช่น truck, sock, box, on, twelfth
2) พยางค์แบบเปิด คือ คำศัพท์ที่ปรากฏสระเพียงหนึ่งเดียวและสระนั้นจะปรากฏอยู่ท้ายพยางค์ เช่น no, a, I, she, he
3) พยางค์ที่ไม่ออกเสียง "e" คือ คำศัพท์ที่ลงท้ายด้วย "e" โดยตัวอักษรก่อนหน้า "e" ต้องเป็นหนึ่งสระตามด้วยหนึ่งพยัญชนะเสมอ เช่น include, ate, slope, these
4) พยางค์ที่มีเสียงสระซ้อน คือ พยางค์ที่มีกลุ่มของสระ 2-3 ตัวเรียงติดกันจะออกเสียง 1 เสียง เสมอ เช่น rain, see, toy, veil
5) พยางค์ที่ออกเสียง R เป็นเสียงสระ คือ พยางค์ที่ลงท้ายด้วยตัวอักษร "r" หรือ "re" รวมทั้งพยางค์ที่มีเสียงสระซ้อนที่ตามด้วยตัวอักษร "r" เช่น car, care, deer
6) พยางค์ที่ประกอบด้วยตัวอักษร L-E คือ พยางค์ที่ประกอบด้วยตัวอักษร "le" เช่น -ble, -cle, -gle และพยางค์ที่พยัญชนะใดๆตามด้วยตัวอักษร "l" เช่น sleep, slope

นอกจากประเภทของพยางค์แล้วการแบ่งคำให้เป็นพยางค์นับเป็นสิ่งจำเป็นในการตัดพยางค์ให้มีความถูกต้องมากยิ่งขึ้น ซึ่งสามารถแบ่งแนวทางในการแบ่งคำได้เป็น 3 แนวทาง ดังนี้

1) แบ่งระหว่าง 2 พยัญชนะที่ปรากฏอยู่คู่กันตรงกลางของคำศัพท์ เช่น hap|pen, bas|ket, let|ter ในกรณีที่เป็นเสียงที่ประกอบด้วยตัวอักษรสองตัวอักษรแต่มีการออกเสียงแค่เพียงตัวอักษรเดียว (Consonant Digraphs) จะไม่มีการแบ่งพยัญชนะ ซึ่งได้แก่ ch, ph, sh, th, wh, ch, ck, sh และ th
2) แบ่งพยางค์ก่อนพยัญชนะ 1 ตัวที่ปรากฏอยู่ตรงกลางของคำเสมอ เช่น o|pen, re|port, i|tem
3) ถ้าหากมี "-le" อยู่ในคำศัพท์ การแบ่งพยางค์จะทำโดยการแบ่งก่อนพยัญชนะที่ตามด้วย "le" เช่น a|ble, rub|ble

เมื่อได้จำนวนพยางค์ที่ถูกต้องแล้ว ขั้นตอนต่อไปเป็นการคำนวณความยาวของพยางค์ต่อคำดังสมการที่ (3.1) จำนวนพยางค์จะถูกนำมาคำนวณหาค่าเฉลี่ยเพื่อหาความยาวของพยางค์ต่อคำทั้งหมดที่มีอยู่ในเนื้อความ

$$F(1) = \frac{\sum_{i=1}^{n} (n_{syl_i})}{W} \qquad (3.1)$$

โดยกำหนดให้

n_{syl_i} คือจำนวนของพยางค์ต่อคำ

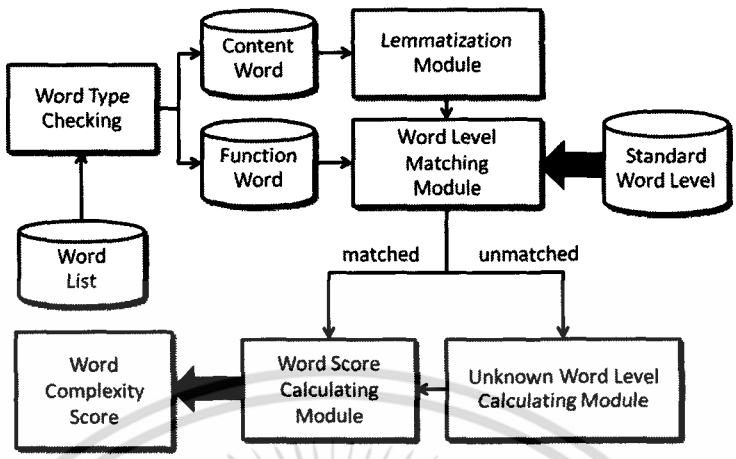
W คือจำนวนของคำทั้งหมดที่ปรากฏภายในเนื้อความ

ผลลัพธ์ของกระบวนการนี้จะได้เป็นค่าคะแนนที่สะท้อนถึงความยาวโดยเฉลี่ยของคำศัพท์ภายในเนื้อความ

3.2.2 การคำนวณความยากของคำศัพท์

การคำนวณความยากของคำศัพท์สำหรับระบบนี้ คือการเปรียบเทียบคำศัพท์ที่ปรากฏอยู่ในเนื้อความสำหรับอ่านกับคำศัพท์ภายในฐานข้อมูลรายการคำศัพท์ที่อ้างอิงกับระดับชั้นที่นักเรียนควรได้เรียนรู้ไว้ ภาพรวมกระบวนการในการคำนวณความยากของคำศัพท์แสดงในรูปที่ 3.3

ในการคำนวณความยากของคำศัพท์นี้ ได้แบ่งประเภทของคำศัพท์ไว้เป็น 2 ประเภทคือ คำบอกความหมาย (Content word) และคำไวยากรณ์ (Function Word) เนื่องจากคำ 2 ประเภทนี้มีนัยยะในการใช้งานทางภาษาที่แตกต่างกันอย่างสิ้นเชิง คำบอกความหมายคือคำที่เปลี่ยนไปตามสถานการณ์และใช้เพื่อสื่อถึงสิ่งของหรือการกระทำอย่างใดอย่างหนึ่ง และเป็นประเภทของคำที่มีเพิ่มขึ้นได้ตลอดเวลา ทว่าคำไวยากรณ์คือคำที่บ่งถึงลักษณะทางไวยากรณ์ของประโยค ไม่มีความสื่อภายในตัวของคำศัพท์นั่นเอง และที่สำคัญเป็นคำที่มีอยู่ในจำนวนจำกัด ดังนั้นการอ่านหรือแปลคำทั้ง 2 ประเภทนี้จึงมีความยากไม่เท่ากันโดยคำไวยากรณ์มักจะปรากฏขึ้นซ้ำๆ ในรูปแบบที่คงที่ มีหลักเกณฑ์แน่นอน จึงมีความง่ายมากกว่า คำบอกความหมายที่จะเปลี่ยนไปตามบริบทหรือสถานการณ์ที่อ้างถึงในแต่ละเนื้อความ



รูปที่ 3.3 กระบวนการการคำนวณความยากของคำศัพท์

ในการคำนวณความยากของคำศัพท์นี้ ได้แบ่งกลุ่มของคำศัพท์ตามหน้าที่ของคำไว้เป็น 2 กลุ่ม คือ 1) คำบอกความหมาย (Content word) และ 2) คำไวยากรณ์ (Function Word) โดยคำทั้ง 2 ประเภทนี้แบ่งได้ชัดเจนตามประเภทของคำ (Part Of Speech: POS) โดยเกณฑ์ในการแบ่งกลุ่มของคำศัพท์ตามหน้าที่ของคำแสดงในตารางที่ 3.1

ตารางที่ 3.1 เกณฑ์ในการแบ่งกลุ่มของคำศัพท์ตามหน้าที่ของคำ

กลุ่มของคำศัพท์หน้าที่	ประเภทของคำ (POS)
คำบอกความหมาย	Noun Verb Adjective Adverb
คำไวยากรณ์	Pronoun Determiner Preposition Conjunction Interjection

ยิ่งไปกว่านั้น คำบอกความหมายที่ปรากฏอยู่ในเนื้อความสำหรับอ่านนั้นมักอยู่ในรูปแบบที่ถูกผัน เช่นการผันคำกริยาตามประธาน (Subject composition) หรือตามกาล (Tense composition) ทำให้ยากต่อการเปรียบเทียบคำศัพท์ ดังนั้นจึงต้องนำคำศัพท์ที่ปรากฏอยู่ในเนื้อความสำหรับอ่านมาสกัดหารูปแท้ (lemma) ก่อน ส่วนคำไวยากรณ์นั้นไม่มีการผันรูปดังนั้นจึงไม่จำเป็นต้องเข้ากระบวนการสกัดหารูปแท้

การสกัดหารูปแท้ (Lemmatization) เป็นการสกัดเอาคำศัพท์ที่ถูกผันออก ซึ่งจะต่างจากการสกัดหารากศัพท์ (Root) ที่จะลบการเติมอุปสรรค (Prefix) หรือปัจจัย (Suffix) ที่เป็นคำเติม (Affixation) ออก ซึ่งจะเปลี่ยนความหมายหรือประเภทของคำศัพท์ไป

การสกัดหารูปแท้ในระบบนี้ ได้ประยุกต์ใช้โปรแกรมเผยแพร่สาธารณะ (Open source tool) ที่ชื่อว่า Morph Adorner โดยเรียกใช้ระบบย่อย English Lemmatizer [12] เพื่อช่วยในการสกัดหารูปแท้ ตัวอย่างการสกัดหารูปแท้แสดงในตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างการสกัดหารูปแท้

คำศัพท์ (Surface word)	รูปแท้ (Lemma)	ประเภทของคำ (POS)
books	book	คำนาม (noun)
feet	foot	คำนาม (noun)
lights, lit	light	คำกริยา (verb)
eats, is eating, has eaten	eat	คำกริยา (verb)
explored, exploring	explore	คำกริยา (verb)
better, best	good	คำวิเศษณ์ (adjective)

เมื่อได้รูปแท้ของคำบอกความหมายแล้ว คำทั้ง 2 ประเภท ถูกนำไปเปรียบเทียบกับฐานข้อมูลรายการคำศัพท์ และใช้สมการ (3.2) ในการคำนวณหาค่าความยากของคำศัพท์

$$F(2)=\frac{\sum_{i=1}^{n}[(Lv_{C_i}*\beta_C*n_C)+(Lv_{f_i}*\beta_f*n_f)]}{\sum_{i=1}^{n}(Lv_{W_i}*W_i)} \quad (3.2)$$

โดยกำหนดให้

Lv คือระดับความยากของคำศัพท์ที่เป็นถูกเปรียบเทียบกับข้อมูลอ้างอิง

f คือชนิดของคำศัพท์ที่เป็นคำไวยากรณ์

c คือชนิดของคำศัพท์ที่เป็นคำบอกความหมาย

β คือสัมประสิทธิ์ของชนิดคำศัพท์

n คือจำนวนของชนิดคำศัพท์ที่ปรากฏในเนื้อความ

W_i คือจำนวนคำศัพท์ทั้งหมดที่ปรากฏในเนื้อความ

หากปรากฏว่า มีคำศัพท์ใดที่อยู่ในเนื้อความแต่ไม่อยู่ในฐานข้อมูลรายการคำศัพท์ ระบบจะจัดการแยกกันระหว่างคำไวยากรณ์และคำบอกความหมาย เนื่องจากคำไวยากรณ์ที่สำคัญและจำเป็นต้อง

เรียนรู้นั้นควรจะถูกจัดเก็บอยู่ในฐานข้อมูลรายการคำไว้แล้วโดยหากเป็นคำที่ไม่ได้จัดเก็บไว้ หมายความว่า เป็นคำประเภท คำอุทาน ที่ไม่ได้มีความหมาย หรือ นัยยะมากนักในการอ่าน ดังนั้นจึงกำหนดให้คำไวยากรณ์ที่ไม่อยู่ในฐานข้อมูลรายการคำศัพท์อยู่ในระดับที่ต่ำสุด คือเทียบเท่ากับคำไวยากรณ์ในกลุ่มประถมต้น ในทางกลับกัน คำบอกความหมายที่ไม่ได้จัดเก็บไว้ในฐานข้อมูลรายการคำศัพท์ มักจะเป็นคำที่เป็นคำศัพท์เฉพาะทางหรือคำที่มีความหมายเฉพาะในเชิงลึก ดังนั้นคำเหล่านี้จึงถูกกำหนดให้เทียบเท่ากับคำบอกความหมายในกลุ่มมัธยมปลายที่เป็นระดับสูงที่สุด

ผลลัพธ์ของกระบวนการนี้จะได้เป็นค่าคะแนนที่สะท้อนถึงความยากของคำศัพท์โดยเฉลี่ยของเนื้อความ

3.2.3 การคำนวณความซับซ้อนของประโยค

ความซับซ้อนของประโยคคือการคำนวณรูปแบบชนิดของประโยค (Sentence type) ในเนื้อความ โดยรูปแบบของประโยคสามารถแบ่งได้เป็น 4 ชนิด ได้แก่

3.2.3.1 ประโยคความเดียว (Simple sentence) คือประโยคที่มีโครงสร้างอย่างง่าย มีประโยคอิสระ (Independent clauses) ปรากฏขึ้นเพียงประโยคเดียว

3.2.3.2 ประโยคความรวม (Compound sentence) คือ ประโยคที่มีตั้งแต่สองอนุประโยคอิสระ (ซึ่งเป็นประโยคความเดียว) ขึ้นไป และแต่ละวลีแยกออกมาเขียนเป็นประโยคความเดียวได้ โดยในประโยคความรวมไม่มี Dependent clauses

3.2.3.3 ประโยคความซ้อน (Complex sentence) คือประโยคที่มีเพียงหนึ่ง อนุประโยคอิสระเท่านั้น แต่จะมีอนุประโยคพึ่งพา (Dependent clause) หนึ่งวลีหรือมากกว่าก็ได้มาเป็นส่วนขยาย

3.2.3.4 ประโยคความซ้ำซ้อน (Compound – complex sentence) คือประโยคที่มีตั้งแต่สองอนุประโยคอิสระขึ้นไป และมีอนุประโยคพึ่งพา ตั้งแต่หนึ่งวลีขึ้นไป กล่าวง่าย ๆ คือมีทั้งประโยคความรวมประโยคความซ้อนปนกันอยู่ ดังในตารางที่ 3.3 แสดงและยกตัวอย่างของรูปแบบประโยคแต่ละชนิด

ตารางที่ 3.3 รูปแบบของประโยคภาษาอังกฤษ

ชนิดของประโยค	ตัวอย่าง
ประโยคความเดียว (simple sentence: S)	- The boy bought those books from the store. - Coral provides good hiding places for fish.
ประโยคความรวม (compound sentence: CPD)	- He sees the recycling truck, and he also sees Janey behind it. - It is really dark in the art gallery, but Harry has a light
ประโยคความซ้อน (complex sentence: CPX)	- Dinosaurs lived on Earth long before there were any people. - When he pushed the last sign into place, the seven lines of light shone into the middle where Henry used to stand on.
ประโยคความซ้ำซ้อน (compound complex sentence: CPDX)	- The man wants the painting, but Harry doesn't have it since the painting has been stolen. - The two friends silently agreed, but their faces showed no fear though the beast rashly ran toward them.

โดยการแบ่งชนิดของประโยค จะใช้คำสำคัญ (Keyword) และการใช้เครื่องหมายวรรคตอนช่วยบ่งตามหลักวากยสัมพันธ์ (Syntax) ภายในประโยค ตามตารางที่ 3.4 และใช้ข้อมูลรายการคำเชื่อมในตารางที่ 3.5 ซึ่งมีคำเชื่อมในประโยคความรวมจำนวน 7 คำ และ คำเชื่อมในประโยคความซ้อนจำนวน 46 คำ

ตารางที่ 3.4 ชนิดของประโยคและเครื่องหมายในการแบ่ง

ชนิดของประโยค	คำสำคัญและการใช้เครื่องหมายวรรคตอน
ประโยคความรวม	พบคำเชื่อมความรวม ตามหลังเครื่องหมายจุลภาค (,) หรือ พบการใช้เครื่องหมาย ";" คั่นประโยค
ประโยคความซ้อน	คำเชื่อมความซ้อนแบบที่ 1 กลางประโยค โดยที่ทั้งประโยคด้านหน้าและด้านหลังเป็นประโยคสมบูรณ์ หรือ พบคำเชื่อมความซ้อนแบบที่ 1 หน้าประโยค โดยมีการใช้เครื่องหมาย ";" คั่นประโยคด้านหน้าและด้านหลัง หรือ พบคำเชื่อมความซ้อนแบบที่ 2 ในประโยคที่ไม่จบด้วยเครื่องหมายคำถาม
ประโยคความซ้ำซ้อน	พบรูปแบบประโยคความรวมและประโยคความซ้อนในประโยคเดียวกัน
ประโยคความเดียว	ไม่ใช่ประโยคความซ้ำ หรือ ประโยคความซ้อน หรือ ประโยคความซ้ำซ้อน

ตารางที่ 3.5 ประเภทคำของคำเชื่อม

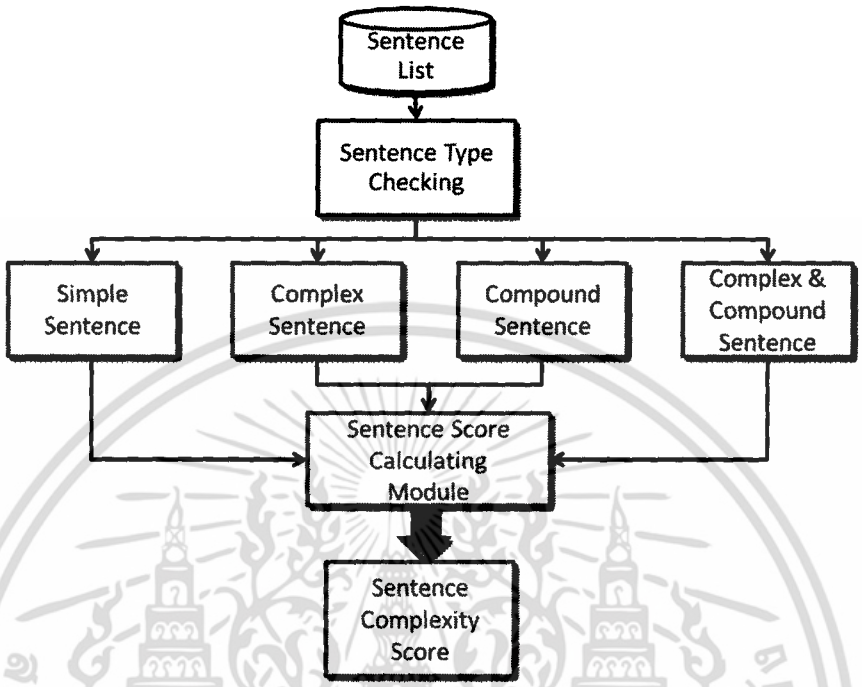
ประเภทคำเชื่อม	ตัวอย่างรายการคำ
คำเชื่อมความรวม	and, or, but, for, yet, nor, so
คำเชื่อมความซ้อนแบบที่ 1	after, although, as, as_if, as_long_as, as_much_as, as_soon_as as_though, because, before, even, even_if, even_though, if, if_only, inasmuch_in_order_that, just_as, lest_ next, now_that once, provided, rather_than, since, so_that, supposing, than, that though, till, unless, until
คำเชื่อมความซ้อนแบบที่ 2	how, what, when, whenever, where, whereas, wherever, whether, which, while, who, whoever, why, that

โดยจากข้อมูลในตารางที่ 3.3 และ ตารางที่ 3.4 ทำให้ได้อัลกอริทึมดังแสดงในรูปที่ 3.4 เพื่อแบ่งชนิดของประโยค

Algorithm for separating sentence type

```
Compound_list = { ",for",",and",",nor",",but",",or",",yet",",so" }
Complex_list = { "after","although","as","as if","as long as","as much as","as soon as",
"as though","because","before","even","even if","even though","how","if","if only",
"inasmuch","in order that","just as","lest","next","now that","once","provided","rather than",
"since","so that","supposing","than","that","though","till","unless","until","what","when",
"whenever","where","whereas","wherever","whether","which","while","who","whoever","why"}
(1)  sentences = getBreakSentence(line);
(2)  For (String sentence : sentences)
(3)  String sentenceTrimed = sentence.trim();
(4)  words = getBreakWord(sentenceTrimed);
(5)       For (String word : words)
(6)            String wordTrimed = word.trim();
(7)            IF(wordTrimed == ",".concat(compound_list)) THEN
(8)                 Found Compound Sentence = true;
(9)                 Compound Sentence Count
(10)           ENDIF
(11)           IF(wordTrimed == complex_list) THEN
(12)                Found Complex Sentence = true;
(13)                Complex Sentence Count
(14)           ENDIF
(15)      ENDLOOPFOR
(16)      IF(Found Compound Or Complex == false) THEN
(17)           Simple Sentence Count
(18)      ENDIF
(19) ENDLOOPFOR
```

รูปที่ 3.4 อัลกอริทึมในการกำหนดชนิดของประโยค



รูปที่ 3.5 กระบวนการการคำนวณความซับซ้อนของประโยค

หลังจากทราบว่า ประโยคต่างๆ ภายในเนื้อความเป็นประโยคชนิดใดบ้าง ระบบจะคำนวณค่าคะแนนความซับซ้อนของประโยคด้วยสมการ (3.3) โดยกำหนดให้มีคะแนนความยากเรียงจากชนิดของประโยคที่ง่ายที่สุดไปยากที่สุด คือ ประโยคความเดียว ประโยคความรวม ประโยคความซ้อน และประโยคความซ้ำซ้อนตามลำดับ

$$F(3)=\frac{\sum_{i=1}^{n}(S*X^{n_x}*P^{n_p})}{Total_{sentence}} \qquad (3.3)$$

โดยกำหนดให้

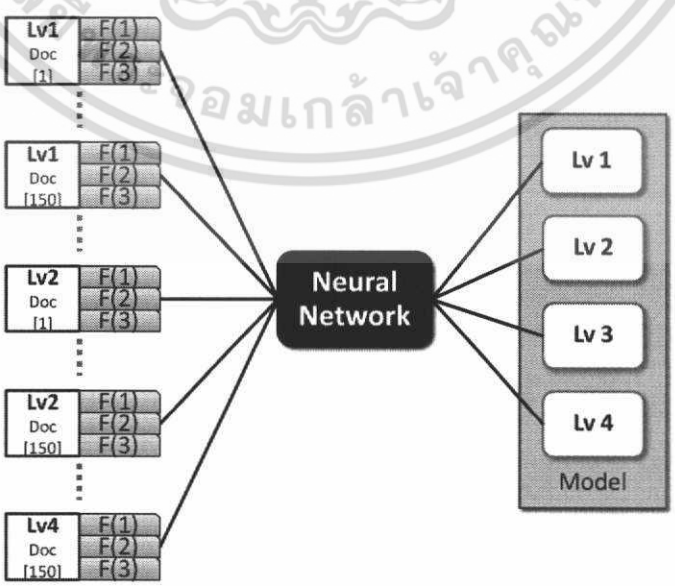
S	คือสัมประสิทธิ์ของประโยคความเดียว
X	คือสัมประสิทธิ์ของประโยคความซ้อน
P	คือสัมประสิทธิ์ประโยคความรวม
n	คือจำนวนของชนิดคำศัพท์ที่ปรากฏในเนื้อความ
$Total_{sentence}$	คือจำนวนประโยคทั้งหมดที่ปรากฏในเนื้อความ

ผลลัพธ์ของกระบวนการนี้จะได้เป็นค่าคะแนนที่สะท้อนถึงความซับซ้อนของประโยคโดยเฉลี่ยของเนื้อความ

3.2.4 การคำนวณค่าความน่าจะเป็นด้วยเครือข่ายประสาทเทียม

กระบวนการการคำนวณค่าความน่าจะเป็นด้วยเครือข่ายประสาทเทียมเป็นกระบวนการสำหรับขั้นตอนการเรียนรู้ เพื่อนำค่าคะแนนจาก 3 กระบวนการก่อนหน้ามาสร้างโมเดลกลุ่มระดับชั้น กระบวนการนี้จะนำค่าคะแนนทั้ง 3 ข้างต้น คือ ค่าคะแนนความยาวของคำ (F(1)) ค่าคะแนนความยากของคำศัพท์ (F(2)) และ ค่าคะแนนความซับซ้อนของประโยค (F(3)) แต่ละตัวมาปรับจูนเป็นค่าความน่าจะเป็น เพื่อสร้างเป็นโมเดลที่บ่งบอกความยากทางภาษาของแต่ละกลุ่มระดับชั้น การทำงานในกระบวนการนี้ได้นำการเรียนรู้ด้วยเครื่องแบบมีผู้สอนคือเครือข่ายประสาทเทียม (Neural network) มาใช้ เครือข่ายประสาทเทียมคือการเรียนรู้ด้วยเครื่อง (Machine learning) ที่สามารถรับข้อมูลรับเข้าเป็นค่าตัวเลข และสร้างเกณฑ์การคัดแบ่งจากค่าตัวเลขเหล่านั้นให้เป็นค่าน้ำหนัก (weight) ที่ช่วยในการประเมินระดับชั้นของเนื้อความ

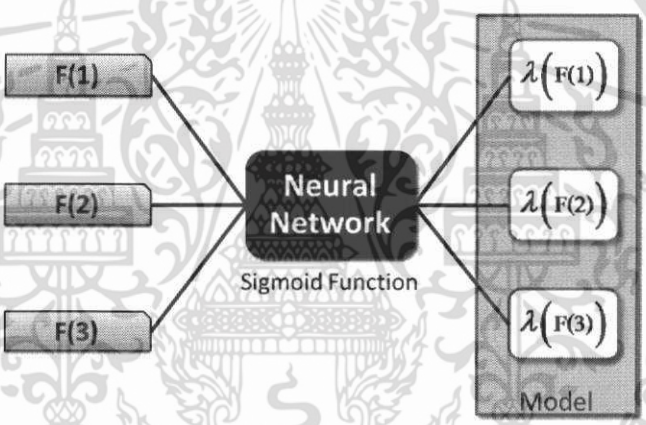
ในงานวิจัยนี้ เนื่องจากมีความซับซ้อนมาก ข้อมูลมีการทับซ้อนกัน ทำให้การสอนโครงข่ายจะมีความซับซ้อนมากขึ้น ดังนั้นจึงใช้การสอนเครือข่ายประสาทเทียมแบบหลายชั้น (Multilayer perception) โดยใช้วิธีการเรียนรู้แบบ Feed forward ใช้ฟังก์ชันถ่ายโอนแบบซิกมอยด์ฟังก์ชั่น (Sigmoid Function) เพื่อปรับเครือข่ายประสาทเทียมให้มีความราบเรียบมากขึ้น



รูปที่ 3.6 การสร้างโมเดลระดับความยากของเนื้อความจากโครงข่ายประสาทเทียม

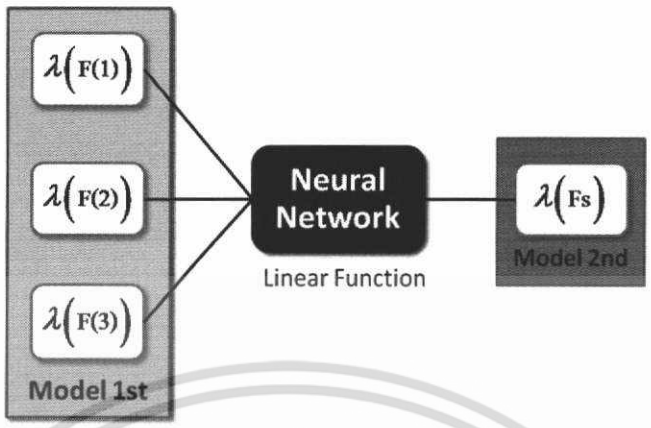
โดยกำหนดค่าอินพุตคือ ค่าคะแนนความยาวของคำ (F(1)) ค่าคะแนนความยากของคำศัพท์ (F(2)) ค่าคะแนนความซับซ้อนของประโยค (F(3)) และเอาต์พุต คือ โมเดลระดับความยากของเนื้อความภาษาอังกฤษของแต่ละกลุ่มระดับชั้น ตามลำดับดังแสดงในรูปที่ 3.6

ทั้งนี้กระบวนการในการสร้างโมเดลของระบบคัดแบ่งระดับเนื้อความโดยใช้เครือข่ายประสาทเทียมนี้ มีการสร้างโมเดลอยู่ 2 กระบวนการ กล่าวคือเมื่อนำค่าคะแนนความยาวของคำ (F(1)) ค่าคะแนนความยากของคำศัพท์ (F(2))และ ค่าคะแนนความซับซ้อนของประโยค (F(3)) แต่ละตัวมาเข้าเครือข่ายประสาทเทียม ผลลัพธ์ที่ได้คือ ค่าความน่าจะเป็นของแต่ละค่า ซึ่งจะแทนด้วยค่า λ โดยจะได้เป็น $\lambda(F(1))$, $\lambda(F(2))$ และ $\lambda(F(3))$ ตามลำดับดังแสดงในรูปที่ 3.7



รูปที่ 3.7 การสร้างโมเดลระดับความยากของเนื้อความจากโครงข่ายประสาทเทียมครั้งที่ 1

หลังจากนั้นจะนำค่าที่ได้ทั้ง 3 ข้างต้นมาสร้างเป็นโมเดลระดับความยากของเนื้อความของแต่ละกลุ่มระดับชั้นโดยนำค่าทั้ง 3 ได้แก่ $\lambda(F(1))$, $\lambda(F(2))$ และ $\lambda(F(3))$ ที่เป็นเวคเตอร์แล้วนำเข้าไปเรียนรู้อีกครั้งด้วยเครือข่ายประสาทเทียมแบบลิเนียร์ (Linear) ดังแสดงในรูปที่ 3.8



รูปที่ 3.8 การสร้างโมเดลระดับความยากของเนื้อความจากโครงข่ายประสาทเทียมครั้งที่ 2

เมื่อได้โมเดลแล้ว โมเดลจะเป็นตัวเทียบกับเนื้อความสำหรับอ่านอื่นๆ และให้ผลลัพธ์เป็นกลุ่มระดับชั้นโดยโมเดลที่ได้จากข้อมูลชุดฝึกสอน (Train Set) จะถูกนำไปเทียบกับข้อมูลชุดทดสอบ (Test Set) เพื่อหาค่าความถูกต้องต่อไป

บทที่ 4
การวางแผนการทดลอง

4.1 เครื่องมือที่ใช้ในการทดลอง

เครื่องมือที่ใช้ในการทดลอง มีคุณสมบัติดังต่อไปนี้

4.1.1 ฮารด์แวร์

หน่วยประมวลผลกลาง (CPU) : Intel Core 2 Duo Processor 2.00 GHz
หน่วยความจำหลัก (RAM) : 2 GB
หน่วยความจำสำรอง (Hard Disk) : 300 GB

4.1.2 ซอฟต์แวร์

ระบบปฏิบัติการ (OS) : Window 7 Professional
โปรแกรมที่ใช้จำลองระบบ : Matlab 7
โปรแกรมที่ใช้ในการจัดการคลังข้อมูล : Edit Plus 3

4.2 วิธีดำเนินการทดลอง

การทดลองในงานชิ้นนี้ จะตรวจสอบความสามารถและความถูกต้องของอัลกอริทึมสำหรับระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านโดยใช้เครือข่ายประสาทเทียม (Passage Grading System Using Neural Network) โดยมุ่งเน้นที่ผลการคัดแบ่งระดับเนื้อความภาษาอังกฤษและผลของแต่ละแง่มุมทางภาษาศาสตร์ทั้ง 3 แบบที่นำมาประยุกต์ใช้ ในการทดลองข้อมูลภายในคลังเนื้อความภาษาอังกฤษสำหรับอ่าน (passage corpus) ถูกแบ่งออกตามระดับชั้นเรียนเป็น 4 ช่วง เพื่อใช้ทั้งในการเรียนรู้ของเครื่อง (machine learning) และการทดสอบความถูกต้อง (accuracy testing) ดังนี้

1) ประถมต้น (ป.1-ป.3)
2) ประถมปลาย (ป.4-ป.6)
3) มัธยมต้น (ม.1-ม.3)
4) มัธยมปลาย (ม.4-ม.6)

ในการคำนวณหาความยากของเนื้อความตามระดับชั้นทั้ง 4 ระดับ แง่มุมทางภาษา 3 ชนิด ได้แก่ ความยาวของพยางค์ ความยากของคำศัพท์ และ ความซับซ้อนของประโยค จะถูกนำมาคำนวณตามสมการที่อธิบายไว้ในบทที่ 3

ดังนั้น ในการทดลองนี้ จึงมีองค์ประกอบหลัก ที่จะใช้ในการทดลองชิ้นนี้ ดังนี้

1) คลังเนื้อความภาษาอังกฤษสำหรับอ่าน แบ่งตาม 4 ช่วงระดับชั้น
2) รายการคำศัพท์เพื่อตรวจสอบความยากของคำ
3) วิธีการในการแบ่งข้อมูลเพื่อทดสอบระบบ

โดยรายละเอียดและรูปแบบข้อมูลของแต่ละองค์ประกอบหลักจะถูกอธิบายในแต่ละหัวข้อย่อยต่อไป

4.2.1 การเตรียมและที่มาของคลังเนื้อความภาษาอังกฤษสำหรับอ่าน

สำหรับการทดลองวัดประสิทธิภาพของระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านจะใช้คลังเนื้อความภาษาอังกฤษ (Passage Corpus) โดยที่เนื้อความภาษาอังกฤษถูกรวบรวมจากแบบการเรียนการสอนของทุกๆชั้นเรียน ตั้งแต่ประถมศึกษาปีที่ 1 – มัธยมศึกษาปีที่ 6 โดยเนื้อความทั้งหมดจะถูกคัดแยกตามแต่ละระดับชั้นเรียนซึ่งกำหนดไว้เป็น 4 กลุ่มระดับ (Grade group) เรียงจากระดับที่ง่ายที่สุดไประดับยากที่สุดตามลำดับ คือ ประถมต้น (ป.1-3) ประถมปลาย (ป.4-6) มัธยมต้น (ม.1-3) มัธยมปลาย (ม.4-6) โดยเนื้อความที่เก็บ เป็นเนื้อความภาษาอังกฤษสำหรับอ่านจากทั้งหนังสือแบบเรียนที่ได้รับการรับรองโดยกระทรวงศึกษาธิการ และหนังสืออ่านนอกเวลาภาษาอังกฤษที่ใช้ในโรงเรียน ตัวอย่างเนื้อความของแต่ละระดับชั้น แสดงในรูปที่ 4.1 ถึงรูปที่ 4.4 ตามลำดับประถมต้น ถึงมัธยมปลาย

> Lions and tigers are big cats. Let's find out about them and other big cats. Tigers are the biggest cats in the world. Tigers have brown fur and dark stripes. Lions are big cats. They live together in families. The mother lion looks after the lion cubs. Leopards like to rest in trees during the day. Leopards are good climbers. Cheetahs are good runners. They are the fastest land animals in the world.

รูปที่ 4.1 แสดงตัวอย่างเนื้อความของประถมต้น

So village people loved to hear stories about Robin Hood. Robin Hood was clever, strong and brave. He loved adventure, and he was the best fighter in England. He took money from rich people and gave it to the poor villagers.

The most famous Robin Hood stories are in this book. They are about beautiful Lady Marian, the greedy Sheriff of Nottingham, good King Richard, and his bad brother, Prince John.

รูปที่ 4.2 แสดงตัวอย่างเนื้อความของประถมปลาย

Egyptians used gold and coloured stones to make jewellery for the rich. When people died, their jewellery and other possessions were buried with them. They believed they could use them in the Afterife. The Pharaoh Tutankhamen wore this amulet to protect him from evil. This piece of jewellery was buried with him.

At the funeral of a Pharaoh, the mummy, in its coffins, was placed on a large sledge. It was pulled across the desert to a tomb. Some tombs were cut deep into desert cliffs. When the sledge reached the River Nile, it was pulled onto a barge, which took it to the other side of the river. Cattle were taken to the tomb and sacrificed. Mourners pulled the mummy on a sledge.

รูปที่ 4.3 แสดงตัวอย่างเนื้อความของมัธยมต้น

Each of his three marriages had followed the same pattern. Using a false name, he had gone on holiday to a place where no one knew him. There he had found a middle-aged, unattractive woman, with some money of her own and no family. He had talked her into marrying him, and she had then agreed to make a will which left him all her money. Both his other wives had been shy too. He was very careful to choose the right type of woman: someone who would not make friends quickly in a new place.Mary, the first of them, had had her deadly 'accident' almost unnoticed, in the bathroom of the house he had rented – a house very like this one, but in the north of England instead of the south. The police had not found anything wrong.

รูปที่ 4.4 แสดงตัวอย่างเนื้อความของมัธยมปลาย

ในการจัดการระเบียบของเนื้อความเบื้องต้น ผู้วิจัยได้คัดแบ่งย่อหน้า (paragraph) ออกเป็นประโยค (sentence) เป็น 1 บรรทัดต่อ 1 ประโยคโดยใช้เครื่องหมายบ่งขอบเขตประโยค เช่น จุดฟูลสตอป (fullstop) เครื่องหมายอัศเจรีย์ (exclamation mark) และ เครื่องหมายคำถาม (question mark)

เป็นตัวบอกใบ้ นอกจากนั้น ผู้วิจัยต้องตรวจสอบการแบ่งคำด้วยช่องว่าง (space) ของเนื้อความด้วยตนเอง เพื่อ ป้องกันความผิดพลาดจากการพิมพ์ เนื่องจากระบบจะแบ่งคำตามช่องว่างระหว่างคำ และจะส่งผลต่อการคำนวณความยากของเนื้อความได้ รายละเอียดในมุมข้อมูลทางสถิติของคลังเนื้อความที่เก็บมา แสดงในตารางที่ 4.1

ตารางที่ 4.1 ข้อมูลทางสถิติของคลังเนื้อความที่ใช้ในการทดลอง

จำนวนและค่าเฉลี่ย	ระดับชั้น			
	ประถมต้น	ประถมปลาย	มัธยมต้น	มัธยมปลาย
จำนวนเนื้อความ	200	200	200	200
ค่าเฉลี่ยจำนวนประโยคต่อเนื้อความ	11.12	30.77	275.93	287.32
ค่าเฉลี่ยจำนวนคำต่อเนื้อความ	42.43	237.83	2050.77	1921.36

4.2.2 การเก็บรายการคำศัพท์เพื่อตรวจสอบความยากของคำ

รายการคำศัพท์ เป็นข้อมูลสำคัญในการวัดระดับความยากของคำศัพท์ภายในเนื้อความ แนวทางในการรวบรวมคำศัพท์มีข้อจำกัดว่า คำศัพท์คำหนึ่งจะมีระดับความยากตามระดับชั้นปีที่คำศัพท์นั้นๆ ได้รับการสอนเป็นครั้งแรก เพียง 1 คำต่อ 1 ระดับเท่านั้น โดยการกำหนดระดับความยากของคำศัพท์อ้างอิงจาก รายการคำศัพท์เพื่อใช้สอนภาษาอังกฤษของกระทรวงศึกษาธิการ และรายการคำศัพท์ในหนังสืออ่านนอกเวลา

ในการวัดระดับความยากของคำศัพท์ ประเด็นสำคัญของความยากคือ ความหมายของคำแต่ละคำ (lexical meaning, semantics) ซึ่งจะผูกพันอยู่กับลักษณะประเภทของคำศัพท์ โดยลักษณะประเภทของคำศัพท์แบ่งได้เป็น 2 ประเภทคือ คำบอกความหมาย (content word) และคำบ่งไวยากรณ์ (function word) เนื่องจากคำประเภทคำบอกความหมาย เป็นคำในกลุ่มเปิด (open class) และมีการใช้งานที่หลากหลาย ไม่มีจำนวนจำกัด เช่นคำนาม (noun) คำกริยา (verb) เป็นต้น ส่วนคำประเภทคำบ่งไวยากรณ์ เป็นคำในกลุ่มปิด (close class) ที่ไม่สื่อความหมายหรือมีความหมายน้อยแต่ ใช้เพื่อบ่งหรือแทรกข้อมูลทางวากยสัมพันธ์ให้กับประโยค และมีจำนวนคำจำกัด เช่น คำสันธาน (conjunction) คำกริยาช่วยบอกกาล (tense expression auxiliary verb) เป็นต้น ดังนั้น การแบ่งคำทั้ง 2 ประเภทออกจากกันภายในรายการคำศัพท์ จึงสามารถช่วยวัดระดับความยากของคำศัพท์ภายในเนื้อความได้อย่างมีประสิทธิภาพ ในกรณีที่เป็นคำศัพท์ที่มี รูปผิว (surface form) เหมือนกัน แต่มีวิธีการใช้ทางไวยากรณ์ต่างกัน หรือมีชนิดของคำ (part of speech) ต่างกัน คำเหล่านั้นจะถูกเก็บแยกกันเป็น 2 คำ ข้อมูลทางสถิติของรายการคำศัพท์เพื่อตรวจสอบความยากของคำ แสดงในตารางที่ 4.2 ตัวอย่างรายการคำศัพท์เพื่อตรวจสอบความยากของคำแสดงในตารางที่ 4.3

ตารางที่ 4.2 สถิติของรายการคำศัพท์เพื่อตรวจสอบความยากของคำที่ใช้ในการทดลอง

ประเภทของคำศัพท์	ระดับชั้น			
	ประถมต้น	ประถมปลาย	มัธยมต้น	มัธยมปลาย
จำนวนคำบอกความหมาย	546	2,030	3,134	3,282
จำนวนคำบ่งไวยากรณ์	30	69	59	32

ตารางที่ 4.3 ตัวอย่างรายการคำศัพท์เพื่อตรวจสอบความยากของคำ

ระดับชั้น	ตัวอย่างคำศัพท์	
	คำบอกความหมาย	คำบ่งไวยากรณ์
ประถมต้น	eat, cat, dog, banana, bee	I, you, he, and, or, but
ประถมปลาย	hot, fine, cabin, beach	for, by, yet, so, else, with
มัธยมต้น	shelter, crew, stun	since, because_of
มัธยมปลาย	heat, election, fabulous	hence, thus

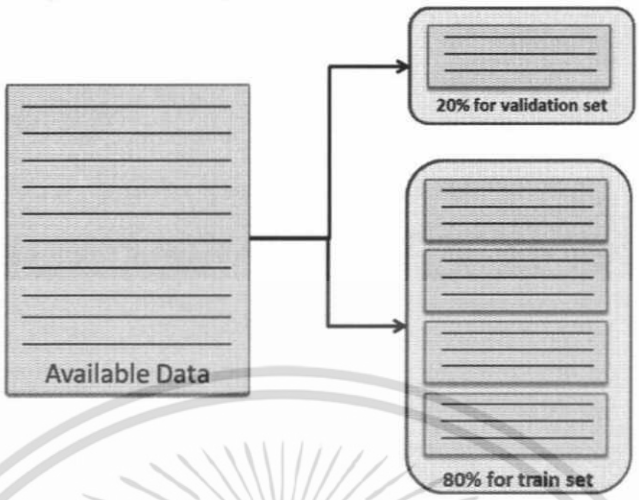
4.2.3 วิธีการในการแบ่งข้อมูลเพื่อทดสอบระบบ

เพื่อตรวจสอบความสามารถของระบบ โดยไม่ต้องใช้มนุษย์ในการให้คะแนนผลการทำงาน ข้อมูลที่ใช้ในการทดลองต้องมีการกำกับระดับชั้นที่ถูกต้องของเนื้อความ แต่เนื่องจากเนื้อความที่มีทั้งหมดภายในคลังเนื้อความมีอยู่จำนวนจำกัด ทำให้ไม่สามารถคัดแบ่งเนื้อความดังกล่าวมาใช้แยกส่วนเป็นข้อมูลทดลองและข้อมูลสำหรับเรียนรู้เพียงอย่างเดียวได้ และการคัดแบ่งเนื้อความแบบเจาะจงนั้นก็ไม่อาจสะท้อนประสิทธิภาพของอัลกอริทึมได้ชัดเจนและไม่เป็นที่ยอมรับเพราะอาจเกิดความลำเอียงในการคัดแบ่งเพื่อให้ผลการทดลองดีขึ้น

ดังนั้นผู้วิจัยได้นำวิธีการแบ่งข้อมูลแบบคลอสวาลิเดชั่น (cross validation) มาประยุกต์ใช้ โดยข้อดีของการวิธีการแบ่งข้อมูลเพื่อทดสอบระบบแบบคลอสวาลิเดชั่นนั้น คือ มีความเหมาะสมและสะท้อนความถูกต้องได้ดีเมื่อทำการทดลองทางสถิติแบบเรียนรู้ด้วยเครื่อง ด้วยปริมาณข้อมูลจำนวนน้อย และเหมาะกับการทดสอบระบบที่อาจจะมีข้อมูลที่ยังไม่เคยเรียนรู้ปรากฏในอนาคต ซึ่งสอดคล้องกับการทดลองชิ้นนี้

วิธีการแบ่งข้อมูลเพื่อทดสอบระบบแบบคลอสวาลิเดชั่น คือแนวทางในการสุ่มแบ่งข้อมูลทั้งหมด ออกเป็นชุดข้อมูล (partition) จำนวนเท่าๆ กัน ตามจำนวนส่วนที่กำหนดไว้ แล้วแบ่ง 1 ชุดข้อมูลเป็น ชุดข้อมูลทดสอบ (validation set) และนำชุดข้อมูลที่เหลือทั้งหมดไปเป็นชุดข้อมูลสอน

ระบบ (train set) ดังแสดงในรูปที่ 4.5 โดยจะวนนำแต่ละส่วนมาเป็นส่วนทดสอบ จนครบ เพื่อจะได้ทดลองกับทุกส่วนข้อมูลอย่างครอบคลุม



รูปที่ 4.5 แสดงการเตรียมชุดข้อมูลเพื่อสอนและทดสอบ

ในการทดลองครั้งนี้ ได้กำหนดให้ แบ่งออกเป็น 5 ส่วน หรือที่เรียกกันว่า 5-โฟลด์ คลอสวาลิเดชั่น (5-fold cross validation) โดยมีการทำงานคือ ให้ชุดข้อมูลแต่ละชุดเขียนแทนด้วย Data1, Data2, ..., Data5 ในครั้งแรกให้ Data2 ถึง Data5 เป็นชุดข้อมูลสอนระบบ และให้ Data1 เป็นชุดข้อมูลทดสอบ ครั้งต่อไปให้ Data1, Data3 ถึง Data5 เป็นข้อมูลสอนระบบ และให้ Data2 เป็นชุดข้อมูลทดสอบ ทำอย่างนี้จนทุกชุดข้อมูลได้เป็นชุดทดสอบ กล่าวคือ ทั้งหมด 5 ครั้งแล้วนำค่าความถูกต้องของการเรียนรู้ทั้ง 5 ครั้งมาหาค่าเฉลี่ย

4.3 ผลการทดลอง

จากการออกแบบระบบคัดแบ่งระดับเนื้อความภาษาอังกฤษสำหรับอ่านสามารถแบ่งผลการทดลองออกเป็น 4 ส่วน ตามแง่มุมทางภาษาศาสตร์ที่นำมาใช้ได้ผลการทดลองดังนี้

4.3.1 ผลการทดลองของการคำนวณความยาวของพยางค์ต่อคำ

การทดลองการคำนวณความยาวของพยางค์ต่อคำเป็นการวิเคราะห์หาค่าเฉลี่ยของพยางค์ที่อยู่ในเนื้อความแต่ละระดับ จากสมการที่ (3.1) ผลลัพธ์ของกระบวนการนี้จะได้เป็นค่าคะแนนที่สะท้อนถึงความยาวโดยเฉลี่ยของคำศัพท์ภายในเนื้อความ ซึ่งจะบ่งบอกถึงความยากของเนื้อความนั้นๆ จากการคำนวณตามสมการที่ (3.1) สถิติของจำนวนพยางค์เรียงตามระดับชั้นเรียนที่ได้ สามารถแสดงในตารางที่ 4.4

ตารางที่ 4.4 คะแนนความยากของจำนวนพยางค์ในแต่ละระดับชั้นเรียน

ระดับชั้นเรียน	จำนวนพยางค์										ค่าเฉลี่ยจำนวนพยางค์
	1		2		3		4		มากกว่า 4		
	คำ	%	คำ	%	คำ	%	คำ	%	คำ	%	
ป.ต้น	5,628	66.32	2,188	25.78	636	7.5	25	0.3	8	0.1	1.42
ป.ปลาย	2,9769	62.58	14,326	30.12	2,949	6.2	380	0.8	142	0.3	1.46
ม.ต้น	229,853	56.04	136,662	33.32	36,339	8.86	4,839	1.18	2,460	0.6	1.57
ม.ปลาย	203,010	52.83	127,616	33.21	38,891	10.12	10,221	2.66	4,534	1.18	1.66

จากตารางที่ 4.4 สังเกตได้ว่าจำนวนพยางค์แต่ละระดับชั้นเรียนมีแนวโน้มเพิ่มขึ้นเมื่อระดับการเรียนการสอนสูงขึ้น

4.3.2 ผลการทดลองของการคำนวณความยากของคำศัพท์

จากการทดลองการคำนวณความยากของคำศัพท์ โดยมีการใช้เกณฑ์คำศัพท์แต่ละระดับชั้นเรียนซึ่งถูกรวบรวมจากหนังสือที่มีการเรียนการสอนจริงจากกระทรวงศึกษาธิการ จากตารางที่ 4.2 สถิติของรายการคำศัพท์จะถูกนำมาใช้เป็นเกณฑ์ในการคำนวณระดับความยากของคำศัพท์ ซึ่งรายการคำศัพท์จะถูกแบ่งออกเป็น 2 ประเภท ได้แก่ คำบอกความหมายและคำบ่งไวยากรณ์ ซึ่งความยากในการอ่านและเรียนรู้ของคำทั้งสองประเภทมีความยากง่ายไม่เท่ากัน ดังนั้นตามสมการที่ (3.2) จึงมีการกำหนดค่าสัมประสิทธิ์แทนค่าความยากของคำในแต่ละประเภท โดยกำหนดให้ คำบอกความหมายมีค่าเท่ากับ 1 และคำบ่งไวยากรณ์ เท่ากับ 0.5 เนื่องคำบ่งไวยากรณ์นั้นเป็นคำที่ง่ายในการเรียนรู้ซึ่งรวมไปถึงคำอุทานในเนื้อความด้วย จึงกำหนดให้ค่าสัมประสิทธิ์ต่ำกว่าคำบอกความหมาย สถิติในการคำนวณความยากของคำศัพท์จะแสดงในตารางที่ 4.5

ตารางที่ 4.5 คะแนนความยากของคำศัพท์ในแต่ละระดับชั้นเรียน

ระดับชั้นเรียน	จำนวนน้อยสุดของความยากคำศัพท์	จำนวนมากสุดของความยากคำศัพท์	ค่าเฉลี่ยความยากของคำศัพท์
ประถมต้น	0.67	1.28	0.97
ประถมปลาย	0.92	1.36	1.10
มัธยมต้น	1.13	2.27	1.31
มัธยมปลาย	1.19	2.83	1.40

4.3.3 ผลการทดลองของการคำนวณความซับซ้อนของประโยค

ความซับซ้อนของประโยคเป็นแง่มุมทางภาษาที่สำคัญอีกแง่มุมหนึ่งในการวัดประสิทธิภาพของเนื้อความภาษาอังกฤษสำหรับอ่าน โดยที่คะแนนความซับซ้อนของประโยคจะถูกคำนวณจากสมการที่ (3.3) ซึ่งการหาคะแนนความซับซ้อนจะมีการคำนวณที่ละประโยคและนำมาหาค่าเฉลี่ยของประโยคทั้งหมดที่อยู่ภายในเนื้อความนั้นๆ โดยกำหนดให้มีคะแนนความยากเรียงจากชนิดของประโยคที่ง่ายที่สุดไปยากที่สุด คือ ประโยคความเดียวกำหนดคะแนน สัมประสิทธิ์เท่ากับ 1 ประโยคความรวมให้กำหนดคะแนนสัมประสิทธิ์เท่ากับ 1.5 และประโยคความซ้อนกำหนดคะแนนสัมประสิทธิ์เท่ากับ 2.0 ตามลำดับ การทดลองคำนวณความซับซ้อนของประโยคจากคะแนนสัมประสิทธิ์ที่กำหนดสามารถหาค่าเฉลี่ยของเนื้อความทั้งหมดโดยแบ่งตามระดับชั้นเรียนได้ผลการดังตารางที่ 4.6

ตารางที่ 4.6 คะแนนความยากของประโยคในแต่ละระดับชั้นเรียน

ระดับชั้นเรียน	ประเภทของประโยค			คะแนนเฉลี่ย
	ประโยคความเดียว	ประโยคความรวม	ประโยคความซ้อน	
ประถมต้น	1947	3	36	1.02
ประถมปลาย	2678	90	712	1.24
มัธยมต้น	30824	3912	9530	1.40
มัธยมปลาย	22008	3148	8286	1.51

จากผลการทดลอง ดังตารางที่ 4.5 สามารถวิเคราะห์ความซับซ้อนของประโยคโดยเปรียบเทียบจากคะแนนสัมประสิทธิ์ที่กำหนด สรุปได้ว่าเนื้อความใดก็ตามที่มีค่าที่มีค่าเฉลี่ยใกล้กับ 1 หมายถึงเนื้อความนั้นมีจำนวนประโยคโดยเฉลี่ยที่มีประโยคความเดียวเป็นจำนวนมากจึงสอดคล้องกับคะแนนเฉลี่ยในระดับประถมต้น ค่าที่เพิ่มขึ้นตามระดับชั้นเรียนบ่งบอกถึงความซับซ้อนที่เพิ่มขึ้น โดยคะแนนที่สูงที่สุดคือมัธยมปลาย เนื่องจากความซับซ้อนของประโยคในเนื้อความส่วนมาก เป็น ประโยคความรวมและประโยคความซ้อน จึงทำให้คะแนนที่ได้อยู่ในเกณฑ์ที่สูงที่สุด

4.3.4 ผลการทดลองการเรียนรู้ของเครื่องด้วยเครือข่ายประสาทเทียม

จากการคำนวณหาคะแนนของสมการในแต่ละแง่มุมทางภาษา ทุกๆเนื้อความจะถูกรวบรวมคะแนนโดยแบ่งออกเป็น 3 ส่วน ได้แก่ คะแนนความยาวของพยางค์ คะแนนความยากของคำศัพท์และคะแนนความซับซ้อนของประโยค ซึ่งข้อมูลของคะแนนในส่วนต่างๆจะถูกนำไปเข้ากระบวนการสำหรับการเรียนรู้ด้วยเครื่องโดยใช้เครือข่ายประสาทเทียมและทดสอบความถูกต้องด้วยวิธีการแบ่งข้อมูลแบบ 5-โฟลด์ คลอสวาลิเดชั่น สำหรับการทดสอบได้มีการกำหนดจำนวนครั้ง

ที่ทดสอบความถูกต้องออกเป็น 5 ส่วน ซึ่งในแต่ละครั้งที่ทดสอบเนื้อความแต่ละระดับชั้นเรียนที่ถูกแบ่งออกเป็น 5 ส่วน จะคิดเป็นร้อยละ 20 ของทุกๆระดับชั้นเรียน ดังแสดงในตารางที่ 4.7

ตารางที่ 4.7 ลำดับเนื้อความที่ใช้ในการทดสอบ

ครั้งที่ทดสอบ	ลำดับเนื้อความ				
	1-40	41-80	81-120	121-160	161-200
Fold-1	ชุดทดสอบ	ชุดการเรียนรู้			
Fold-2	ชุดการเรียนรู้	ชุดทดสอบ	ชุดการเรียนรู้		
Fold-3	ชุดการเรียนรู้		ชุดทดสอบ	ชุดการเรียนรู้	
Fold-4	ชุดการเรียนรู้			ชุดทดสอบ	ชุดการเรียนรู้
Fold-5	ชุดการเรียนรู้				ชุดทดสอบ

จากการทดลองการระบุระดับความยากของเนื้อความภาษาอังกฤษสำหรับอ่านโดยวิธีการเรียนรู้ด้วยเครื่องโดยใช้เครือข่ายประสาทเทียม ในการทดสอบครั้งที่ 3 มีจำนวนของเนื้อความที่ถูกต้องมากที่สุดในการระบุระดับความยากคือ 141 เนื้อความจากเนื้อความทั้งหมด 160 เนื้อความ และจำนวนของเนื้อความที่ระบุระดับความยากถูกน้อยที่สุดคือ 132 เนื้อความ จากเนื้อความทั้งหมด 160 เนื้อความ ในการทดสอบครั้งที่ 4 และจากการทดสอบความถูกต้องทั้ง 5 ครั้ง การระบุระดับความยากของเนื้อความโดยใช้เครือข่ายประสาทเทียมสามารถระบุระดับเนื้อความได้ถูกต้องทั้งหมด 683 เนื้อความ จากทั้งหมด 800 เนื้อความ โดยคิดเป็นร้อยละได้ 85.38 ดังตารางที่ 4.8

ตารางที่ 4.8 ผลการทดลองการระบุระดับของเนื้อความโดยใช้เครือข่ายประสาทเทียม

ระดับชั้นเรียน	จำนวนเนื้อความที่ถูกต้องในแต่ละครั้ง					รวม	
	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	จำนวน	ร้อยละ
ประถมต้น	35/40	35/40	36/40	34/40	35/40	175/200	87.5
ประถมปลาย	33/40	34/40	36/40	33/40	34/40	170/200	85
มัธยมต้น	33/40	35/40	34/40	32/40	34/40	168/200	84
มัธยมปลาย	32/40	36/40	35/40	33/40	34/40	170/200	85
รวม	133/160	140/160	141/160	132/160	137/160	683/800	85.38

จากการระบุระดับเนื้อความโดยใช้เครือข่ายประสาทเทียมได้ผลลัพธ์ความถูกต้องร้อยละ 85.38 จากตารางที่ 4.8 จำนวนความถูกต้องในการระบุเนื้อความในแต่ละระดับชั้นเรียนที่มากที่สุดได้แก่ ชั้นประถมต้น โดยมีจำนวนที่เครือข่ายประสาทเทียมระบุได้ถูกต้อง คือ 175 เนื้อความจาก

ทั้งหมด 200 เนื้อความ โดยหากนำแต่ละแง่มุมทางภาษามาทดลองหาระดับความยากของเนื้อความโดยใช้วิธีการเรียนรู้ด้วยเครื่องจะได้ผลลัพธ์ ดังตารางที่ 4.9

ตารางที่ 4.9 ค่าความถูกต้องแต่ละแง่มุมทางภาษา

ระดับชั้นเรียน	ค่าความถูกต้อง		
	พยางค์(ร้อยละ)	คำศัพท์(ร้อยละ)	ประโยค(ร้อยละ)
ประถมต้น	66.19	81.29	77.70
ประถมปลาย	64.75	81.29	79.86
มัธยมต้น	67.15	66.42	84.67
มัธยมปลาย	66.42	65.67	87.31
เฉลี่ย	66.12	73.77	82.33

4.4 ปัจจัยและประเด็นที่ส่งผลต่อความสามารถของอัลกอริทึม

ปัจจัยและประเด็นที่ส่งผลต่อความสามารถของอัลกอริทึม แบ่งออกเป็น 3 ส่วน คือ 1. ปัจจัยจากส่วนข้อมูลที่ใช้ในการเรียนรู้ 2. ปัจจัยจากแง่มุมทางภาษาศาสตร์ที่นำมาใช้ และ 3. ปัจจัยจากค่าสัมประสิทธิ์ที่กำหนดให้ในการคำนวณระดับความยากของแต่ละแง่มุมภาษา

4.4.1 ปัจจัยจากส่วนข้อมูลที่ใช้ในการเรียนรู้

ในการเก็บเนื้อความเพื่อมาเป็นข้อมูลที่ใช้ในการเรียนรู้ พบว่าปัญหาและผลกระทบในการนำเนื้อความมาใช้เพื่อเรียนรู้เกิดจาก มาตรฐานและคุณภาพของเนื้อความที่นำมาใช้เป็นหลัก ในงานวิจัยชิ้นนี้ เนื้อความถูกเก็บมาจาก 2 แหล่งหลักๆ คือ 1.เนื้อความภาษาอังกฤษสำหรับอ่านจากทั้งหนังสือแบบเรียนที่ได้รับการรับรองโดยกระทรวงศึกษาธิการ และ 2.หนังสืออ่านนอกเวลาภาษาอังกฤษที่ใช้ในโรงเรียน ซึ่งเนื้อความทั้ง 2 ชนิดนี้มีความแตกต่างกันมากในเชิงความยากตามหลักภาษาศาสตร์ และพบว่า ข้อความจากหนังสืออ่านนอกเวลาภาษาอังกฤษที่ใช้ในโรงเรียน มีความยากกว่า ทั้งด้านการใช้คำศัพท์ และรูปแบบประโยค ซึ่งความต่างของความยากที่ได้กล่าวมานั้น ส่งผลกระทบ ต่อการคำนวณระดับความยากของเนื้อความของอัลกอริทึมและก่อให้เกิดความเหลื่อมล้ำของความยากในแต่ละระดับ จากการทดลองหาค่าเฉลี่ยเฉพาะเนื้อความของระดับประถมต้นและประถมปลาย หากแยกเนื้อความทั้ง 2 ประเภทออกจากกันและนำไปใช้คำนวณระดับความยากตามแง่มุมทางภาษาศาสตร์ทั้ง 3 ที่ได้กำหนดมานั้น จะได้ผลลัพธ์ดังแสดงในตารางที่ 4.10

ตารางที่ 4.10 เปรียบเทียบคะแนนระหว่างเนื้อความจากแบบเรียนและหนังสือนอกเวลา

ระดับชั้นเรียน	ประเภทเนื้อความ	พยางค์	คำศัพท์	ประโยค
ประถมต้น	เนื้อความจากแบบเรียน	1.36	0.83	1.01
	เนื้อความจากหนังสือนอกเวลา	1.46	1.11	1.03
ประถมปลาย	เนื้อความจากแบบเรียน	1.40	1.05	1.21
	เนื้อความจากหนังสือนอกเวลา	1.48	1.16	1.27

จากข้อมูลในตาราง พบว่า ความยากของเนื้อความจากหนังสือนอกเวลาในระดับประถมต้นนั้น มีค่าคะแนนความยากสูงกว่าเนื้อความจากแบบเรียนสำหรับชั้นประถมปลาย และสาเหตุนี้ทำให้ช่องห่างของระดับความยากแคบและเหลื่อมล้ำกันในบางจุด ซึ่งจะส่งผลให้การตัดสินระดับความยากของเนื้อความผิดได้

อย่างไรก็ตาม เพื่อการรองรับการใช้งานจริงของผู้ใช้ การตัดเนื้อความส่วนใดส่วนหนึ่งออกเป็นเรื่องที่ไม่สมควร เนื่องจากผู้ใช้มีความเป็นไปได้ที่จะนำเนื้อความที่มาจากทั้งเนื้อความจากแบบเรียนและเนื้อความจากหนังสือนอกเวลามาตรวจหาระดับความยาก และหากตัดส่วนใดส่วนหนึ่งออก ก็จะเป็นการจำกัดการใช้งานของผู้ใช้ไป ดังนั้นผู้วิจัยขอเสนอว่า การพัฒนางานนี้ต่อ ควรมีการแบ่งประเภทเนื้อความไว้ชัดเจนก่อนการเรียนรู้ และควรเพิ่มทางเลือกให้ผู้ใช้เลือกก่อนว่าเนื้อความที่จะนำมาทดสอบมาจากแหล่งข้อมูลประเภทไหน แล้วจึงนำไปคำนวณหาระดับความยาก ซึ่งจะเพิ่มอัตราความแม่นยำได้มากขึ้นและสมเหตุสมผลยิ่งขึ้นอีกด้วย

4.4.2 ปัจจัยจากแง่มุมทางภาษาศาสตร์ที่นำมาใช้

ประเด็นถัดมาคือการจัดการข้อมูลเมื่อนำไปใช้คำนวณความยากด้วยแง่มุมทางภาษาศาสตร์ที่กำหนดไว้ โดยส่วนที่ทำให้ประสิทธิภาพตกลงมีอยู่ 3 จุดด้วยกันคือ 1) คำจำนวนพยางค์น้อยแต่ยากกว่าคำจำนวนพยางค์มาก ในคำนวณความยากจากจำนวนพยางค์ 2) การไม่สามารถจำแนกคำหลายความหมายได้ (Polysemous word) ในการหาความยากของคำศัพท์ และ 3) คำที่ไม่พบในรายการคำศัพท์

1) ความยากจากจำนวนพยางค์

คำจำนวนพยางค์น้อยแต่ยากกว่าคำจำนวนพยางค์มากนั้น ส่งผลให้การคำนวณความยากจากจำนวนพยางค์ เกิดความผิดเพี้ยนไปจากสมมติฐานที่คาดไว้ เนื่องจาก สมมติฐานได้คาดไว้ว่า คำที่มีจำนวนพยางค์มากกว่า ส่วนมาก น่าจะยากกว่า คำที่มีจำนวนพยางค์น้อย โดยความเป็นจริงนั้น พบว่า ภายในเนื้อความระดับมัธยมปลาย มีคำที่มีจำนวนพยางค์ 1 ถึง 2 พยางค์ ไม่ต่างกับเนื้อความระดับชั้นอื่นๆ หากแต่คำเหล่านั้น เป็นคำศัพท์ที่ไม่ค่อยนิยมใช้และมีความหมายเฉพาะ เช่นคำว่า “woe” “agog” “foe” เป็นต้น ดังนั้นผลการคำนวณความยากจากจำนวนพยางค์จึงไม่ส่งผลลัพธ์ได้

อย่างที่คาดหวังไว้และมีประสิทธิภาพต่ำในภาพรวมของการคัดแบ่งความยากของเนื้อความสำหรับอ่าน นอกจากนั้น แง่มุมทางภาษาศาสตร์ยังมีข้อจำกัดจากระบบนับพยางค์ที่นำมาประยุกต์ใช้อีกด้วย เนื่องจากระบบดังกล่าวในปัจจุบันไม่สามารถนับจำนวนพยางค์ที่ถูกต้องได้ทั้งหมด เช่นการนับพยางค์ระหว่างคำว่า "resume" ที่อ่านว่า เรซูเม้ (3 พยางค์) และ รีซูม (2 พยางค์)ได้ โดยทุกๆ ครั้งที่เจอคำนี้ระบบจะนับแค่เพียง 2 พยางค์ทุกครั้ง ทำให้การนับพยางค์มีความผิดพลาดเกิดขึ้น ดังนั้นเพื่อพัฒนางานต่อในอนาคต การนับจำนวนพยางค์จึงไม่ควรเป็นแง่มุมหลักในการใช้คัดแบ่งความยากของเนื้อความ

2) ความยากจากคำหลายความหมาย

คำหลายความหมายเป็นอีกหนึ่งปัจจัยที่ทำให้การคัดแบ่งความยากของเนื้อความสำหรับอ่านมีประสิทธิภาพต่ำลง คำหลายความหมายในภาษาอังกฤษมีจำนวนมากและ พบว่าความหมายที่แตกต่างกันมีระดับความยากที่แตกต่างกันด้วย เช่นคำว่า "plant" คำนี้ หากแปลว่า "ต้นไม้ (คำนาม)" เป็นคำที่จะได้รับการสอนในระดับประถมชั้นปีที่3 และถูกจัดเก็บอยู่ในรายการคำศัพท์ชั้นประถมต้น หากแปลว่า "การปลูกต้นไม้ (คำกริยา)" เป็นคำที่จะได้รับการสอนในระดับประถมชั้นปีที่5 และถูกจัดเก็บอยู่ในรายการคำศัพท์ชั้นประถมปลาย หากแปลว่า "โรงงาน (คำนาม)" เป็นคำที่จะได้รับการสอนในระดับมัธยมชั้นปีที่2 และถูกจัดเก็บอยู่ในรายการคำศัพท์ชั้นมัธยมต้น อย่างไรก็ตาม ภายในอัลกอริทึม ไม่สามารถแบ่งความยากจากคำที่เขียนเหมือนกันทุกประการได้ ทำให้ทุกครั้งที่เจอคำศัพท์คำนี้ จะถูกกำหนดเป็นคำที่มีความยากระดับประถมต้น ซึ่งไม่สะท้อนความยากที่แท้จริงของคำศัพท์ ในการแก้ปัญหานี้ในอนาคต ผู้วิจัยเสนอให้นำ ระบบจัดการความกำกวมทางด้านความหมาย (word sense disambiguation system) มาใช้ เพื่อแบ่งแยกความหมายของคำ เพื่อให้การคำนวณความยากตามแง่มุมความยากของคำศัพท์ มีประสิทธิผลเพิ่มขึ้น

อย่างไรก็ตาม ในปัจจุบัน ระบบจัดการความกำกวมทางด้านความหมายนั้น ยังไม่เป็นที่แพร่หลายและให้ความถูกต้องที่ต่ำ การจัดการปัญหานี้ขั้นต้น สามารถทำได้ด้วยการนำ ระบบกำหนดลักษณะของคำ (Part-of-speech tagger) มาใช้เพื่อแบ่งคำที่มีลักษณะของคำ (Part-of-speech) ต่างกันมาใช้เพื่อแบ่งคำหลายความหมายที่มีลักษณะของคำต่างกันก่อนได้ เช่นในกรณีของคำว่า "plant" นี้ อย่างน้อย ก็จะสามารถแบ่งระดับประถมต้นและประถมปลายออกจากกันได้

3) คำที่ไม่พบในรายการคำศัพท์

รายการคำศัพท์เพื่อตรวจสอบความยากของคำ (vocabulary word list) เป็นส่วนสำคัญในการวัดระดับความยากของคำศัพท์ (vocabulary usage) ภายในเนื้อความ อย่างไรก็ตามคำในรายการคำศัพท์ ไม่อาจครอบคลุมทุกคำในเนื้อความได้ เช่นคำประเภท ชื่อเฉพาะ (proper name) คำบอกเวลาเฉพาะ (date and time entity) หรือ คำในระดับที่สูงกว่า เช่น คำศัพท์ระดับมหาวิทยาลัย และคำศัพท์เฉพาะสาขาวิชาคำเหล่านี้ในอัลกอริทึมปัจจุบัน ผู้วิจัยได้แก้ปัญหาโดยการ กำหนดให้ คำที่ไม่มีในรายการที่เป็นคำบอกความหมาย (unknown content word) มีความยากเท่ากับ คำศัพท์บอก

ความหมาย ระดับมัธยมปลายเพื่อให้ได้คะแนนความยากสูงสุด ส่วนคำที่ไม่มีในรายการที่เป็นคำไวยากรณ์ (unknown function word) มีความยากเท่ากับคำไวยากรณ์ ระดับประถมต้นเพื่อให้ได้คะแนนความยากต่ำที่สุด เพราะผู้วิจัยสมมติฐานว่าคำบอกความหมายที่ไม่มีในรายการ ควรจะเป็นคำเฉพาะในระดับที่สูงกว่าหรือชื่อเฉพาะเท่านั้น ในทางกลับกัน คำไวยากรณ์ที่เป็นคำในวงจำกัด (close class) ที่ไม่มีในรายการจะเป็นเพียงคำอุทาน (interjection) ที่ไม่มีนัยยะสำคัญในเนื้อความ

อย่างไรก็ดี จากการทดลองพบว่า ภายในเนื้อความสำหรับอ่าน มีคำประเภท ชื่อเฉพาะ (proper name) และ คำบอกเวลาเฉพาะ (date and time entity) มากเป็นพิเศษ แม้กระทั่งในระดับประถมต้น ทำให้ส่งผลต่อการวัดระดับความยากในแง่มุมความหมายของคำศัพท์ ดังนั้นคำเหล่านี้จึงควรจะได้รับการจัดการแยกเป็นพิเศษ เช่นการประยุกต์ใช้ระบบรู้จำเอนทิตี้ระบุนาม (name entity recognition) และคัดแยกออกจากคำศัพท์ที่ต้องนำไปหาความยาก เพื่อไม่ให้เกิดความกำกวมในการคิดค่าความยากต่อไป

ยิ่งไปกว่านั้น ยังมีคำศัพท์อีกจำนวนมากซึ่งเป็นคำที่ไม่มีอยู่ในรายการคำศัพท์ แต่ไม่ได้เป็นคำศัพท์ในระดับที่สูงกว่า เช่น คำศัพท์ระดับมหาวิทยาลัยและคำศัพท์เฉพาะสาขาวิชา หากแต่รายการคำศัพท์ไม่ครอบคลุม เนื่องจากรายการคำศัพท์ที่เก็บมาจาก รายการคำศัพท์เพื่อใช้สอนภาษาอังกฤษของกระทรวงศึกษาธิการ และรายการคำศัพท์ในหนังสืออ่านนอกเวลา ไม่ได้มีทุกคำศัพท์ที่ใช้ในการเรียนการสอนภาษาอังกฤษทั้งหมด ทำให้ปรากฏคำศัพท์ที่ไม่รู้จักอีกมาก นอกจากนี้ ยังมีปัจจัยในเชิงการสะกดคำที่ต่างกันของรายการคำศัพท์ และคำที่ปรากฏจริงในเนื้อความด้วย เพราะคำในรายการคำศัพท์นั้น เป็นคำศัพท์ในรูปแบบการสะกดตามรูปแบบอเมริกา (American English) แต่ในเนื้อความจริง บางส่วนเป็นคำศัพท์ในรูปแบบการสะกดตามรูปแบบอังกฤษ (British English) ทำให้ไม่สามารถหาคำที่ตรงกันได้และปรากฏเป็นคำที่ไม่รู้จักไปแทน เช่นคำว่า “color - colour” “jewelry – jewellery” หรือ “recognize – recognise” เป็นต้น ผู้วิจัยเสนอว่า การเก็บรวบรวมคำศัพท์เป็นรายการคำศัพท์นั้น ควรเก็บจากเนื้อความจริงในแต่ละระดับชั้นเพื่อให้เกิดความครอบคลุม มากกว่า การใช้รายการคำศัพท์ที่กำหนดเอง เพราะไม่อาจจะสะท้อนการใช้คำศัพท์ได้ทั้งหมดและมีข้อจำกัดในรูปแบบการสะกด

4.4.3 ปัจจัยจากค่าสัมประสิทธิ์ที่กำหนดให้

ในการทดลอง ผู้วิจัยได้กำหนดค่าสัมประสิทธิ์ในการให้คะแนนแก่แต่ละประเภทของประโยคและคำศัพท์เอง เช่น กำหนดให้ ค่าสัมประสิทธิ์ ตัวคูณ ของประโยคความซ้ำ และความซ้อน เท่ากับ 1.5 และ 2.0 ตามลำดับ หรือ กำหนดให้ ค่าสัมประสิทธิ์ ตัวคูณ ของคำบอกความหมายและคำไวยากรณ์ เท่ากับ 1.0 และ 0.5 ตามลำดับ ค่าเหล่านี้เป็นค่าที่กำหนดจากมุมมองความรู้ทางภาษาศาสตร์ที่ได้ไปศึกษาและสอบถามจากนักภาษาศาสตร์มาโดยตรง อย่างไรก็ตามค่าเหล่านี้

ส่งผลโดยตรงต่อผลการคำนวณขออัลกอริทึม กล่าวคือ หากค่าสัมประสิทธิ์เหล่านี้เปลี่ยน ผลการทดลองก็จะเปลี่ยนตามไปด้วย

ในความเป็นจริง ค่าสัมประสิทธิ์เหล่านี้ ไม่มีแหล่งที่มาและไม่ได้เกิดจากการคำนวณทางสถิติใดๆ โดยระบบ หากแต่เป็นการกำหนดจากผู้เชี่ยวชาญ ซึ่งไม่อาจครอบคลุมปรากฏการณ์ของภาษาในทุกกรณี ดังนั้น ในการพัฒนางานชิ้นนี้ต่อ ผู้วิจัยจึงเสนอให้นำค่าส่วนนี้มากำหนดโดยอัตโนมัติด้วย การคำนวณทางด้านสถิติด้วยเพื่อให้ ค่าเหล่านี้ สามารถปรับเปลี่ยนได้ตามปรากฏการณ์ของภาษา และสมเหตุสมผลมากกว่าในเชิงการจัดการทางสถิติแบบการเรียนรู้ด้วยเครื่อง

บทที่ 5
วิเคราะห์ผลการทดลองและสรุป

เนื้อหาในบทนี้จะกล่าวสรุปผลการทดลองเปรียบเทียบกับสมมติฐานที่ตั้งไว้ และวิเคราะห์ผลการทดลอง พร้อมทั้งเสนอแนวทางในการพัฒนางานให้มีประสิทธิภาพดียิ่งขึ้น และแนวทางการแก้ปัญหารวมทั้งวิธีการแก้ไขข้อจำกัดของงานชิ้นนี้

5.1 วิเคราะห์ผลการทดลองและสรุปผล

วิทยานิพนธ์ฉบับนี้ ได้นำเสนออัลกอริทึมสำหรับคัดแบ่งระดับความยากของเนื้อความสำหรับอ่านภาษาอังกฤษ (English reading passage) โดยใช้สมการในการคำนวณค่าคะแนนความซับซ้อนจากแง่มุมทางภาษาศาสตร์ 3 ชนิด (linguistic features)ได้แก่จำนวนพยางค์ (syllable length) ความยากของคำศัพท์ (vocabulary complexity) และ ความซับซ้อนของประโยค (sentence complexity) และใช้การเรียนรู้ด้วยเครื่อง (machine learning) แบบเครือข่ายประสาทเทียม (neural network) เพื่อสร้างโมเดลทางสถิติของระดับความยาก (statistical model of passage level) จากค่าคะแนนความซับซ้อนดังกล่าวเพื่อใช้วัดระดับความยากของเนื้อความสำหรับอ่านภาษาอังกฤษ

ในการทดลอง ผู้วิจัยได้เก็บรวบรวมคลังเนื้อความสำหรับอ่านภาษาอังกฤษ(English passage corpus) จำนวน 800 เนื้อความ โดยแบ่งเป็น 4 ช่วงระดับชั้นคือ ประถมต้น ประถมปลาย มัธยมต้น และมัธยมปลาย และแบ่งเป็นช่วงระดับชั้นละ 200 เนื้อความ ที่มาของเนื้อความมาจาก 2 แหล่งคือ 1) เนื้อความภาษาอังกฤษสำหรับอ่านจากทั้งหนังสือแบบเรียนที่ได้รับการรับรองโดยกระทรวงศึกษาธิการ และ 2) หนังสืออ่านนอกเวลาภาษาอังกฤษที่ใช้ภายในโรงเรียน

เพื่อวัดประสิทธิภาพของอัลกอริทึมจึงเลือกใช้วิธีการตรวจสอบความถูกต้องแบบ 5-โฟลด์ คลอสวาลิเดชั่น (5-fold cross validation) โดยการแบ่งข้อมูลทั้งหมดออกเป็น 5 ส่วน นำ 4 ส่วนไปเป็นชุดข้อมูลสอนระบบ และ 1 ส่วนเป็นชุดทดสอบ จากนั้นจะสลับสับเปลี่ยนชุดทดสอบวนไปจนครบทั้ง 5 ส่วน กล่าวคือ ข้อมูลทั้งหมดจะได้ใช้ทดสอบวัดความถูกต้องโดยเป็นทั้ง ชุดข้อมูลสอน และ ชุดทดสอบอย่างครบถ้วน

จากผลการทดลอง ปรากฏว่า อัลกอริทึมสำหรับคัดแบ่งระดับความยากของเนื้อความสำหรับอ่านภาษาอังกฤษ ให้ความถูกต้องในการระบุความยากอยู่ที่ประมาณร้อยละ 85 นอกจากนั้นหากมองแยกแต่ละแง่มุมทางภาษาศาสตร์ พบว่า ประสิทธิภาพของแต่ละแง่มุมไม่เท่ากัน โดยความซับซ้อนของประโยค ความยากของคำศัพท์ และจำนวนพยางค์ เป็นแง่มุมที่ให้ผลความถูกต้องได้มีประสิทธิภาพสูงสุดไปน้อยสุด ตามลำดับ

หากวิเคราะห์แยกตามช่วงระดับชั้นเรียน จากตารางที่ 4.9 พบว่า ความยากของคำศัพท์ มีประสิทธิภาพในการระบุความยากของเนื้อความดีที่สุด สำหรับการคัดแบ่งช่วงระดับชั้นประถมต้นและประถมปลาย ส่วน ความซับซ้อนของประโยค แสดงศักยภาพได้ดีในการคัดแบ่งช่วงระดับชั้นมัธยม ในทางกลับกัน จำนวนพยางค์ ที่ให้ผลลัพธ์ประมาณร้อยละ 66 ทั้งในประถมและมัธยมเป็นแง่มุมทางภาษาศาสตร์ที่ระบุความยากของเนื้อความได้ไม่น่าพอใจ

ยิ่งไปกว่านั้น จากการทดลองพบว่า อัลกอริทึมและวิธีการมีข้อจำกัดอยู่พอสมควรดังได้กล่าวไว้ในหัวข้อ 4.4 ได้แก่

1) ความแตกต่างกันอย่างมากของเนื้อความต่างประเภทกัน
2) คำศัพท์ที่มีรูปผิวอย่างเดียวกันแต่มีระดับความยากต่างกัน
3) คำศัพท์ที่ไม่อยู่ในรายการคำศัพท์
4) คำศัพท์ที่มีการสะกดหลายแบบ
5) การกำหนดค่าสัมประสิทธิ์โดยนักภาษาศาสตร์

ข้อจำกัดและปัจจัยเหล่านี้ ส่งผลให้ค่าความถูกต้องของคำทำนายลดลง ซึ่งปัญหาเหล่านี้ยังไม่ได้รับการแก้ไขในงานปัจจุบัน

5.2 ข้อเสนอแนะและแนวทางการพัฒนางานวิจัย

แนวทางการพัฒนาอัลกอริทึมนี้ต่อไปในอนาคต สิ่งแรกที่ควรให้ความสำคัญคือการแก้ไขจุดบกพร่องหรือข้อจำกัดที่ส่งผลให้ค่าความถูกต้องของการระบุความยากของเนื้อความลดลง ซึ่งแต่ละจุดบกพร่องหรือข้อจำกัดมีวิธีการแก้ไขหรือจัดการไม่เหมือนกัน

ปัญหาแรกคือ ความแตกต่างกันอย่างมากของเนื้อความต่างประเภทกัน ปัญหานี้เกิดจากการรวบรวมเนื้อความเป็น คลังเนื้อความ โดยมิได้ จัดแบ่งประเภทของเนื้อความไว้ก่อน ดังนั้น หากเก็บรวบรวมแบ่งประเภทไว้ และทำทางเลือกให้ผู้ใช้งาน ได้เลือกประเภทให้ตรงกับเนื้อความของตนเอง อัลกอริทึมก็จะสามารถให้ค่าในการระบุระดับเนื้อความมีความแม่นยำได้สูงขึ้น นอกจากนั้นยังสามารถนำเอกลักษณ์เฉพาะของแต่ละประเภทเนื้อความมาเพิ่มเป็นแง่มุมทางภาษาศาสตร์อื่นๆ ได้อีกด้วย เช่น เนื้อความจากหนังสือนอกเวลา มีแนวโน้มว่า จะมีรูปแบบประโยคเป็นแบบ “คำรายงานจากคำพูดตนเอง (direct speech)” มากกว่าเนื้อความจากแบบเรียนที่มักจะเป็นภาษาเขียนแบบเรียงความ ดังนั้นการแบ่งประเภทเนื้อความจึงเป็นแนวทางที่ควรนำมาประยุกต์ใช้กับอัลกอริทึมนี้เป็นอย่างมาก

ปัญหาถัดมา เป็นปัญหาในการจัดการข้อมูลให้รองรับการคำนวณหาค่าคะแนนความยากจากสมการในการหาความยากของคำศัพท์ ภาษาอังกฤษนั้นเป็นภาษาที่มีวงคำศัพท์ไม่มาก โดยคำศัพท์จำนวนหนึ่ง เป็นคำที่มีรูปผิว (surface) เดียวกัน แต่ใช้ได้หลายตำแหน่งการใช้งานและ

หลากหลายความหมาย ซึ่งการตรวจสอบลักษณะการใช้งานโดยตรวจสอบเพียงรายการคำศัพท์อย่างเดียวไม่สามารถบ่งบอกความหมายได้ ดังนั้นจึงควรหาอัลกอริทึมหรือระบบที่ช่วยแบ่งความหมายของคำ เช่น ระบบบอกความหมายจากคำรอบข้าง (word sense disambiguation from collocation) เป็นต้น

นอกจากนั้น ปัญหาคำนามชื่อเฉพาะ ก็ส่งผลให้ค่าคะแนนสูงขึ้นเพราะเป็นคำที่ไม่มีในรายการคำศัพท์ การแก้ไขที่เสนอแนะคือการนำระบบรู้จำชื่อ (name recognition) มาใช้และตัดชื่อเฉพาะเหล่านี้ออก จากการคิดค่าความยากเพราะคำเหล่านี้ ไม่ได้มีนัยยะสำคัญในมุมมองความหมายในการอ่านเนื้อความ

ส่วนปัญหาคำศัพท์ที่มีไม่เพียงพอในรายการคำศัพท์และปัญหาคำศัพท์ที่มีการสะกดหลายรูปแบบนั้น แนวทางการแก้ไขที่เสนอแนะคือ การนำคำศัพท์จากภายในเนื้อความที่อยู่ในชุดเรียนรู้เข้ามาผนวกกับรายการคำศัพท์ด้วย โดยอาจจะนำค่าทางสถิติของการเกิดของคำนั้นมาช่วยในการร่วมทำนายระดับชั้น ซึ่งจะส่งผลให้วงคำศัพท์ในรายการคำมีความครอบคลุม

ในการพัฒนาศักยภาพของงานชิ้นนี้ ผู้วิจัยเสนอแนะว่า ควรนำแง่มุมทางภาษาศาสตร์อื่นมาใช้ร่วมด้วย เนื่องจาก แง่มุมทั้ง 3 ไม่อาจสะท้อนความยากของเนื้อความได้ทั้งหมด ตัวอย่างแง่มุมทางภาษาศาสตร์อื่นๆ อาทิเช่น การใช้สำนวน การอ้างอิงสรรพนามที่กล่าวมาแล้ว และจำนวนประโยค เป็นต้น จะช่วยเพิ่มมุมมองในการคัดแบ่งระดับชั้นได้ละเอียดมากขึ้น

นอกจากนี้ หากพัฒนาอัลกอริทึมนี้ต่อ ร่วมกับการสร้างโมเดลตามความสนใจในการเลือกเนื้อความของนักเรียนจะสามารถพัฒนาให้เป็นระบบช่วยเหลือการเรียนการสอนในห้องเรียนได้ เพราะ อัลกอริทึมนี้จะช่วยลดเวลาในการตรวจเนื้อความของอาจารย์ และสามารถจำกัดและค้นหาเนื้อความที่ตรงกับจุดด้อยของนักเรียนที่ต้องการพัฒนาความสามารถในการอ่านได้อย่างตรงจุด

เอกสารอ้างอิง

[1] Flesch R (1948). "A new readability yardstick". *Journal of Applied Psychology 32*, pp. 221–233.

[2] Kincaid JP, Braby R, Mears J (1988). "Electronic authoring and delivery of technical information". *Journal of Instructional Development 11*, pp. 8–13.

[3] Kincaid JP, Braby R, Wulfeck WH II (1983). "Computer aids for editing tests". *Educational Technology 23*, pp. 29–33.

[4] Braby R, Kincaid JP, Scott P, McDaniel W (1982). "Illustrated formats to teach procedures". *IEEE Transactions on Professional Communications 25*, pp. 61–66.

[5] Kincaid JP, Aagard JA, O'Hara JW, Cottrell LK (1981). "Computer Readability Editing System". *IEEE Transactions on Professional Communications 24*, pp. 38–42.

[6] McClure G (1987). "Readability formulas: Useful or useless. (an interview with J. Peter Kincaid.)". *IEEE Transactions on Professional Communications 30*: pp. 12–15.

[7] Gunning, R. (1952). The technique of clear writing. New York, NY: McGraw-Hill International Book Co.

[8] http://simbon.madpage.com/Fog/fog.cgi

[9] McLaughlin, G. Harry (May 1969). "SMOG Grading — a New Readability Formula". *Journal of Reading 12 (8)*, pp. 639–646. Retrieved 2008-09-20.

[10] http://www.harrymclaughlin.com/SMOG.htm

[11] Tom M. Mitchell (1997) Book: Machine Learning p.2

[12] สมิช บัตรเจริญ. การเปรียบเทียบผลการพยากรณ์ปริมาณเลขหมายของชุมสายโทรศัพท์ด้วยเทคนิคการถดถอยกับเทคนิคเครือข่ายประสาทเทียม. สารนิพนธ์ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ , 2547.

[13] วรุณย์พันธุ์ สุขสมมโน. การปรับจูนตัวเองของตัวควบคุมพีไอดีโดยใช้ข่ายงานระบบประสาท.กรุงเทพฯ, จุฬาลงกรณ์มหาวิทยาลัย, 2542.

[14] เหรียญพงษ์ จุฬาวิเศษกุล. การวิเคราะห์ความสั่นสะเทือนโดยใช้นิวรัลเน็ทเวิร์ค. วิทยานิพนธ์ ปริญญาวิศวกรรมศาสตร์มหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ ,จุฬาลงกรณ์มหาวิทยาลัย , 2536.

[13] Fausett, L., 1994, "Fundamentals of Neural Networks: Architectures, Algorithms, and Application," Prentice-Hall, USA.

[14] Hornik, K., Stinchcombe, M., White, H., 1989, "Multilayer feedforward networks are Universal Approximators," Neural Networks 2, pp. 359-366.

[15] http://morphadorner.northwestern.edu/

ภาคผนวก ก
ผลงานที่ได้รับการตีพิมพ์

KICSS 2010

Proceedings of the Fifth International Conference on Knowledge, Information and Creativity Support Systems

KICSS 2010

November 25-27, 2010
Chiang Mai, Thailand

Thanaruk Theeramunkong, Virach Sornlertlamvanich,
and Susumu Kunifuji (Eds.)

ISBN 978-974-466-504-1

Passage Grading System Using Supervised Learning

Wasan Na Chai† **Nualsawat Hiransakolwong†**

Taneth Ruangrajitpakorn‡ **Thepchai Supnithi‡**

† Computer Science department,
King Mongkut's Institute of Technology Ladkrabang, Thailand
‡Human Language Technology Laboratory,
National Electronic and Computer Technology Center, Thailand
wasannachai@gmail.com, khnualsa@kmitl.ac.th
{taneth.ruangrajitpakorn, thepchai.supnithi}@nectec.or.th

Abstract

In this paper, we propose an approach to check a level of English reading passage by applying supervised learning to create a level characteristic model from corpus. The system calculates from three features; word, syllable and sentence complexity. The system does not require manual criteria to grade a passage but automatically grade by comparing to a model. The output of the system shows a level of passage based on four groups categorised by Thai academic school level. With the purposed system, 84.95% of passages are correctly graded.

Keywords: grading system, English reading passage, readability, supervised learning

1 Background

For nations that English language is not their native language, English learning is a necessary subject since English is the dominant language for global communication. English becomes more difficult to be effectively taught for people whose native language belongs to a different language typology such as Thai. Based on the statistics of ETS TOEFL score [1], Thai people gains 493 points of a paper-based test in average. English learning in Thailand; therefore, particularly needs more improvement and attention.

Among four English learning skills (writing, reading, speaking and listening), reading could be preliminarily concentrated since it helps learners to gently acquire new vocabularies and to instance English correct structures. However, reading is the most wearisome part for learners because all reading passages are usually selected by teachers, not learners. Hence, learners lost their motivation and attention on reading the English passage because it is not met students' interest. However, if teachers let students choose a reading passage on their own will, the chosen passage will be difficult to be controlled in terms of an appropriateness of passage level and content.

Several grading systems have been implemented to fulfil such need, for example Flesch reading easy formula, Kincaid formula, SMOG-Grading and Fox index. The Flesch reading easy formula [2] has been developed by Flesch in 1948 and it is based on school text covering grade 3 to 12. Unfortunately, it has not been updated for a decade. The Kincaid Formula [3] has been developed for grading Navy training manuals. It is accountable in technical document grading because it is based on adult training manuals rather than school book text. The SMOG-Grading [4] is a tool for grading English texts. It has been developed by McLaughlin in 1969. Its result is a school grade. The Fog index [5] has been developed by Robert Gunning. It especially concerns a proper name issue and handles it separately.

All of above mentioned systems have their own advantage, and they all compute a readability score based on syllable, word, and sentence amount. Their scores are graded by manually constructed criteria. However, they all employed their own manually constructed criteria which apparently suit for only native English people. Thus, their output is not practical to students who learn English as foreign language (EFL). In this paper, we focus on EFL of Thai students in academic school.

The rest of this paper is organised as follow. Section 2 proposes a supervised learning grading system with its three complexity scores. Section 3 describes experiment setting and grading result. Section 4 discusses over the re-

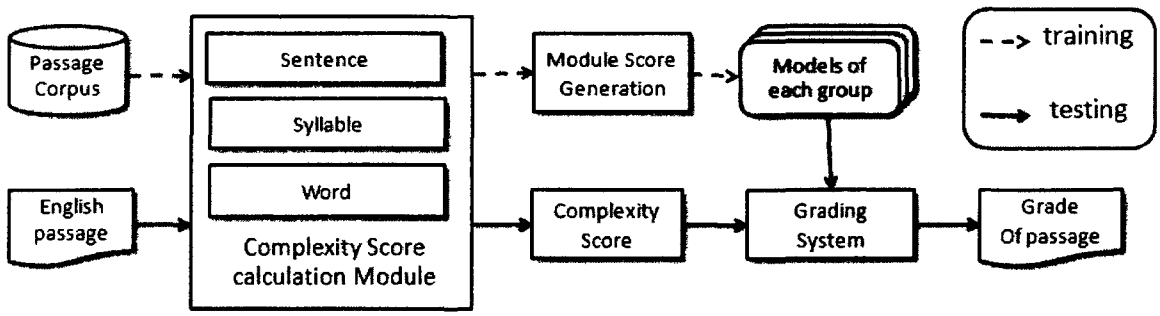


Figure 1. An overview of the supervised passage grading system

sult. Section 5 explains the further usage of the system with user model strategy. Last, section 6 gives a conclusion and future work.

2 Passage grading system

We propose a method for an English passage grading by using a supervised learning to create a model from an academic passage corpus to compare with student selected passage [6]. In both training a model from corpus and grading a passage, three complexity scores are exploited to signify a level of a passage. The three complexity scores are a complexity of syllables, vocabulary level and sentence type.

For models generation, training corpus collected from academic English reading passages are split into four grade groups based on their belonged level. The four groups are junior primary school (grade 1-3), primary school (grade 4-6), junior high school (grade 7-9) and high school (grade 10-12). With the complexity scores of each grade group, a characteristic model is generated.

Once characteristic models are obtained, they are used as a reference to grade the student's selected passage to an appropriate level that the passage belongs to. A system overview of supervised learning grading system is illustrated in Figure 1.

2.1 Syllable complexity score module

A length of syllable is a measure of a lexical difficulty. Based on our survey of the training corpus, we found that the more syllable a word contains, the more difficult level it could be. An average length of syllables in a context significantly reflects difficulty of student's readability. A statistic of syllable length in the training corpus is illustrated in Table 1.

Table 1. Percentages of syllable length

Grade	Percentage of syllable length					Avg.
	1	2	3	4	4<	
junior primary school	66.32	25.78	7.5	0.3	0.1	1.42
primary school	62.58	30.12	6.2	0.8	0.3	1.46
junior high school	56.04	33.32	8.86	1.18	0.6	1.57
high school	52.83	33.21	10.12	2.66	1.18	1.67

To calculate a complexity score of syllable, we obtain (1).

$$F1 = \frac{\Sigma_{i=1}^{n}(n_{syl_i})}{W} \quad (1)$$

Where syl_i is the number of syllable of word i^{th} and W is the total number of words in a passage.

2.2 Vocabulary complexity score module

A vocabulary in a reading passage becomes an essential feature for students' readability. In general, students learn vocabularies by a given list of assigned vocabularies in a school textbook based on their academic level. Thus, all vocabularies provided in textbooks that were approved by the Ministry of Education of Thailand are chosen as a reference word list for vocabulary level. The purpose of reference list is to be matched words in a reading passage. Words in a passage, however, are usually in inflected form which cannot be directly matched to the word list therefore they must be transformed into a lemma form[1] by applying

[1] A lemma form refers to the canonical form, dictionary form, or citation form of a word, e.g., in English, eat, eats, ate, eaten and eating are forms of the same lexeme, with run as the lemma. It is different from a word stem which all affixes are removed.

Morpha [7][8] before they are calculated for a complexity score. In the system, words are classified into two groups, content word[2] and function word[3] because of their different significance. They are calculated with a different value.

For the unknown word issue, a function word and a content word are separately handled. Based on reference word list, practical function words are entirely collected, but there are some missing rare function words such as exclamation words which are seldom used and have very little meaning. They are thus assigned equally to the lowest level of function word. Otherwise, missing content words from reference word list reckon with a greater level of vocabularies because the reference word list is gathered with all vocabularies that students should learn in school. The missing content words normally are beyond school vocabularies. An unknown content word, therefore, is given as same as the highest grade of content word.

With above criteria, a vocabulary complexity score is computed by (2).

$$F2 = \frac{\Sigma_{i=1}^{n}[(Lv_{c_i} \cdot \beta_c \cdot n_c) + (Lv_{f_i} \cdot \beta_f \cdot n_f)]}{\Sigma_{i=1}^{n}(Lv_{w_i} \cdot W_i)} \quad (2)$$

where Lv refers to a level of a word in reference list, c indicates a content word, f is a function word, β is a parameter, n is a frequency, w refers to any kind of word and W_i is a frequency of i^{th} word.

A process of vocabulary complexity score calculation is shown in Figure 2.

2.3 Sentence type complexity score module

Basically, a sentence type in English falls into four types which are simple sentence (S), compound sentence (CP), complex sentence (CX) and compound-complex sentence (CPX). For the last three types, a sentence contains more than a single independent clause and they can cause a difficulty to students on reading a context. Apparently, the more clauses a sentence has, the more difficulty it gains.

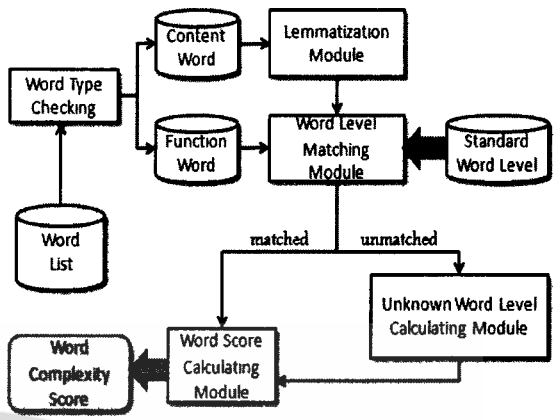


Figure 2. A vocabulary complexity score calculation process

For the system, each sentence in a passage must be chunked into clause(s) to analyse a sentence type. The complexity score of sentence type is computed based on an amount of complexness of clause(s) by using (3).

$$F3 = \frac{\Sigma_{i=1}^{n}(S \cdot P^{n_P} \cdot X^{n_X})}{Total_{sentence}} \quad (3)$$

where S refers to a simple sentence, X indicates a complex sentence and P is a compound sentence. N_x is a number of a recursion of a complex sentence and N_p is a recursive number of compound sentence. The detail and examples of sentence types within the passage corpus are shown in Table 2.

2.4 Model score module

After the three complexity scores are obtained, artificial neural network [9][10] is applied to weight them for representing characteristics model of passages. We obtained (4) to compute a model score of each grade group.

$$Model = \lambda_1 F1 + \lambda_2 F2 + \lambda_3 F3 \quad (4)$$

where λ indicate a weighted variable of complexity score gained by the artificial neural network and F1, F2 and F3 refer to syllable, vocabulary and sentence type complexity score respectively.

2.5 Grading module

To grade a level of a passage, a test passage is submitted to all complexity score modules to calculate its complexity. With the models, the sum of those scores is checked which model it belongs to. The result of this module is a level

[2] Content words are words that have a stable lexical meaning, such as noun, verb and adjective.

[3] Function words are words that have little lexical meaning or have ambiguous meaning, but instead serve to express grammatical relationships with other words within a sentence such as pronoun, conjunction, preposition and interjection.

of the student selected passage.

Table 2. Examples of four sentence types

Type	Example
S	- The boy bought that book from the store yesterday. - Coral provides good hiding places for fish and other creatures. - I will do the same with the third child tonight.
CP	- He sees the recycling truck, and he also sees Janey behind it. - It is really dark in the art gallery, but Harry has a light. - The wind was not so strong now and there was a fullmoon.
CX	- Dinosaurs lived on Earth long before there were any people. - When he pushed the last sign into place, the seven lines of light shone into the middle where Henry used to stand on. - Although he was badly frightened, Stephen decided to go to Mr. Abney's library.
CPX	- The man wants the painting, but Harry doesn't have it since the painting has been stolen. - The two friends silently agreed, but their faces showed no fear though the beast rashly ran toward them. - Stephen apparently knew that it wasn't a joke, and he was really worried about the issue.

3 Experiment

3.1 Experiment setting

In this paper, we centre on a readability of Thai learner in EFL. English reading passages from textbooks approved by the Ministry of Education of Thailand and supplementary readings assigned in school are selected as a corpus. They are split into four groups by academic grade; junior primary school (grade 1-3), primary school (grade 4-6), junior high school (grade 7-9) and high school (grade 10-12). Statistic of details in gathered reading passage corpus of each group is shown in Table 3.

An English vocabulary list for Thai learners was only garnered from the Ministry of Education of Thailand. Table 4 shows a statistic of the reference word list of each grade group for Thai passage grading. The reference word list with level and word type is exemplified in Table 5.

Table 3. a statistic of reading passage corpus

grade group	passage amount	average sent/psge	average word/ psge
junior primary school	53	32.85	338
primary school	66	212.8	1271.5
junior high school	98	244.62	1771.17
high school	133	287.32	1961.75
total	350	777.59	5342.42

Table 4. a statistic of reference word list

grade group	content word	function word	total
junior primary school	546	30	576
primary school	2,030	69	2,099
junior high school	3,134	59	3,193
high school	3,282	32	3,314
total	8,992	190	9,182

Table 5. Examples of reference word list

Vocabulary	Word type	Grade level
eat,cat,dog,banana,bee	content	junior primary school
I,you,he,and,or,but	function	
hot,fine,cabin,beach	content	primary school
for,by,yet,so,else,with	function	
shelter,crew,stun	content	junior high school
since,because_of	function	
heat,election,fabulous	content	high school
hence,thus	function	

In the experiment, the parameters for calculating vocabulary complexity are initially set as 1.5 for a content word and 1.0 for a function word because a content word which is an open set of words is generally more difficult than a function word which is a close set. For the parameters of the sentence complexity score, simple sentence, compound sentence and complex sentence are set as 1.0, 1.2 and 1.4 respectively because a complex sentence is concerned as the most difficult sentence type, and a simple sentence is accounted that it is the easiest sentence type. In the experiment, we applied a feed-forward neural network [10] to weight the complexity scores. A multi-layer perceptron and a

(a) Lions and tigers are big cats. Let's find out about them and other big cats. Tigers are the biggest cats in the world. Tigers have brown fur and dark stripes. Lions are big cats. They live together in families. The mother lion looks after the lion cubs. Leopards like to rest in trees during the day. Leopards are good climbers. Cheetahs are good runners. They are the fastest land animals in the world.

(b) Egyptians used gold and coloured stones to make jewellery for the rich. When people died, their jewellery and other possessions were buried with them. They believed they could use them in the Afterife. The Pharaoh Tutankhamen wore this amulet to protect him from evil. This piece of jewellery was buried with him.

At the funeral of a Pharaoh, the mummy, in its coffins, was placed on a large sledge. It was pulled across the desert to a tomb. Some tombs were cut deep into desert cliffs. When the sledge reached the River Nile, it was pulled onto a barge, which took it to the other side of the river. Cattle were taken to the tomb and sacrificed. Mourners pulled the mummy on a sledge.

Figure 3. Examples of an English reading passage

sigmoid function were also exploited to manage continuous derivative, which allows it to be used in backpropagation.

To exemplify a calculation of the three complexity scores, the passage examples of a junior primary school and a primary school are shown in Figure 3. In the Figure 3a, the passage from a junior primary school contains 74 words which are 49 content words and 25 function words. There are few content words that belong to higher level in this passage such as "stripes", "cubs" and "climbers". For syllable aspect, the total and average amount of syllables and the longest syllable are 94, 1.27 and three, respectively. Because of the lowest level, all eleven sentences in the passage are a simple sentence. The example in Figure 3a obtains the complexity scores from syllable, vocabulary, and sentence type as 1.27, 1.16 and 1.0, respectively

From the passage of a primary school in Figure 3b, it totally contains 124 words which are split into 66 content words and 58 function words. The function words in the passage are such as "and", "to" "the", "in", "with", and "across". The total and average amount of syllables and the longest syllable are 177, 1.427 and four, respectively. In addition, the passage contains seven simple sentences and five complex sentences. There is a complex sentence that contains two complex clauses within it as "When the sledge reached the River Nile, it was pulled onto a barge, which took it to the other side of the river." With the system calculation, the complexity scores from syllable, vocabulary, and sentence type are 1.43, 1.26 and 1.10 respectively.

3.2 Experiment result

In our experiment, 5-folds cross validation [11] is applied. 80% of each grade group in the corpus are randomly selected for training passage grade models and the rest is for testing. A statistic of a reading passage corpus is previously mentioned in Table 3. A result of grading system is shown in Table 6.

Table 6. An accuracy result of grading system

Group	fold-1	fold-2	fold-3	fold-4	fold-5	avg.
1	78.33	81.49	82.06	79.61	84.06	81.11
2	82.9	77.64	74.71	79.12	80.31	78.94
3	84.03	83.36	84.03	90.23	88.16	85.96
4	93.04	95.31	96.33	91.78	92.59	93.81
all	84.58	84.45	84.28	85.19	86.28	84.95

From the results, the system gains impressive accuracy as 84.95%. For each of the complexity score, they are separately shown in Table 7, Table 8 and Table 9 to show accuracy by each single feature.

Table 7. An accuracy result of F1

Group	fold-1	fold-2	fold-3	fold-4	fold-5
1	80.46	82.36	74	79.61	84.36
2	81.98	71.04	74.95	76.87	80.31
3	82.76	82.71	84.03	90.44	87.93
4	92.41	95.3	96.49	91.28	93.58
all	84.4	82.85	82.37	84.55	86.55

Table 8. An accuracy result of F2

Group	fold-1	fold-2	fold-3	fold-4	fold-5
1	71.33	75.89	85.64	81.11	80.06
2	79.37	78.94	70.91	80.04	76.8
3	79.87	84.96	80.76	86.76	82.76
4	89.84	92.02	93.63	88.04	90.89
all	80.1	82.95	82.74	83.99	82.63

Table 9. An accuracy result of F3

Group	fold-1	fold-2	fold-3	fold-4	fold-5
1	83.16	86.19	87.14	78.31	87.66
2	87.45	82.8	78.11	82.67	83.96
3	89.93	81.31	87.03	93.76	93.87
4	96.77	98.31	98.83	95.78	96.79
all	89.33	87.15	87.78	87.63	90.57

4 Discussion

From the result, the major error occurs in the vocabulary complexity module. We found that there are many unknown words within the passages. Since they come from different sources, they are explicitly different and this results incompatibility between reference word list and vocabulary in the reading passages. Since we handle unknown content words as the highest level vocabulary, the passages contained many unknown words are resulted to higher school level than they actually are.

Moreover, we discover that a polysemous word (a word with multiple meanings) cannot be handled in the system effectively. This causes problem with vocabulary complexity score calculation since the matched surface words are always recognised as the lowest level in the reference word list. This problem leads to grade a passage to a lower grade group than it should be.

With above mentioned issues, we found that they reflex the vocabulary complexity score to gain the lowest accuracy among other. If the word sense disambiguation is applied accordingly, the system would perform better in term of vocabulary recognition and overall accuracy.

The other errors in the syllable and sentence type complexity are from the similarity of difficulty of the passage among different levels. From our observation, there are not much different in terms of amount of sentence types and an average number of syllables in the reading passage between grade group 1 and grade group 2 as well as between grade group 3 and grade group 4. In the other hand, the explicit differences are the length of the passage and the frequency of an idiom and a proverb. To improve accuracy, these two points are needed to be focused constantly and should be added for other features.

5 Grading system with user model

For further usage of the grading system, it can be applied in e-learning system as a core engine to grade a passage for students. In EFL e-learning, students' model is a representation of their information which is a key to draw their attention and directly improve their weaker skills among other.

The model gathers many aspects of personal information of students. The personal information can be split into three parts: 1) personal profile; 2) interested subject and activity; and 3) test result. The first part is a personal profile which details about student's name, gender, current grade, etc. The interested subject and activity includes their desired occupation, hobby, sport, favourite food and beverage, favourite song and film genre, etc. The last part is a test result. The test is quizzed for finding their strong and weak fluency of English skills. The test includes vocabulary quiz, passage reading test, syntactic structure quiz, pronunciation test and essay writing test.

With the information, we can develop a passage retriever which selects an appropriate passage to fulfil their interest and match for students' weak skills efficiently for improving such skills. Their weaker skills obtained from the test are matched with the complexity scores of the grading system. There are two different aspects on using the system: to improve students' skills and to assist students to enjoy their reading.

To improve students' skills, the system can select a challenging passage that is graded to higher level. The higher level passages are challenged them to improve their weak skills by selecting the appropriate complexity score. For example, students whose syntactical structure is weak are rationally provided with the passage with high sentence complexity score. In the other hand, students who are not good at pronunciation and vocabulary are given with the passage which is classified for high vocabulary

complexity score.

To help students to enjoy their reading, the system can be applied to select the suitable passage matched to students' level and weak skills. For example, student whose skill is weak in vocabularies is assigned with a reading passage of his/her grade that contain lower vocabulary and syllable complexity scores for them to read a passage with joy and understanding. For student who needs more practice in English syntactic structure, a passage with higher sentence type complexity score is given.

In summary, the grading system applied in EFL e-learning based on user model can recommend a passage that matches students' interest and their English skills. This will improve their English learning with student attention and motivation.

6 Conclusion and future work

In this paper, we present a passage grading system which uses supervised learning approach to create a passage model of each grade group. Four grade groups are divided according to a school grade; junior primary school (grade 1-3), primary school (grade 4-6), junior high school (grade 7-9) and high school (grade 10-12). A passage model includes a complexity score of syllable, vocabulary and sentence type by using neural network to adjust the appropriateness of its complexity scores. The system then matches an input passage with a model and returns a grade that the input passage belongs to. In the experiment, reading passages used in English classed for Thai students are collected as a corpus to generate a passage model. The system obtains an impressive result of 84.95% accuracy when testing with 5-fold cross validation.

In the future, we plan to apply neural network in vocabulary complexity score calculation module for tuning parameters of content word and function word since they are manually set as 1.0 and 0.5 respectively in the current work. Moreover, we plan to apply this system to e-learning application for an automatic passage retrieval based on students' model mentioned in Section 5. In addition, topic selection will be applied to match users' interest topic. Word sense disambiguation will be plugged to the system for handling a polysemous word issue. We also plan to add another feature such as frequency of idiom and proverb appeared in the passage, a length of a reading passage, or a number of specific vocabularies in different domain. Lastly, another supervised learning approach will be exploited to compare and find the best learning for supervised grading system.

References

[1] ETS (2010), TOEFL Test and Score Data Summary for TOEFL Computer-Based and Paper-Based Tests: 2009 Test Year Data.

[2] R. Flesch, 1984. "A New Readability Yardstick," Journal of Applied Psychology, 32, 221-233.

[3] L. Si and J. Callan., 2001. A Statistical Model for Scientific Readability. In Proceedings of the tenth international conference on Information and knowledge management.

[4] H. McLaughlin, 1969. "SMOG Grading — a New Readability Formula". Journal of Reading 12 (8): 639–646.

[5] S. Fuller, C. Horlen, R. Cisneros, and T. Merz, 2007. "Pharmacy Students' Reading Ability and the Readability of Required Reading Materials". American Journal of Pharmaceutical Education 2007; 71 (6) Article 111.

[6] W. Na Chai, T. Ruangrajitpakorn, T. Supnithi., 2010. A Statistical Approach on Automatic Passage Level Checking Framework for English Learner. To be appeared In Proceedings of ICCE 2010.

[7] G. Minnen, J. Carroll and D. Pearce., 2000. 'Robust, applied morphological generation'. In Proceedings of the 1st International Natural Language Generation Conference, Mitzpe Ramon, Israel. 201-208.

[8] Morphological and Orthographic Tools for English (morpha): available online at http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/morph.html

[9] J. A. Anderson, 1995, Introduction to Neural Networks, Cambridge MA: Press.

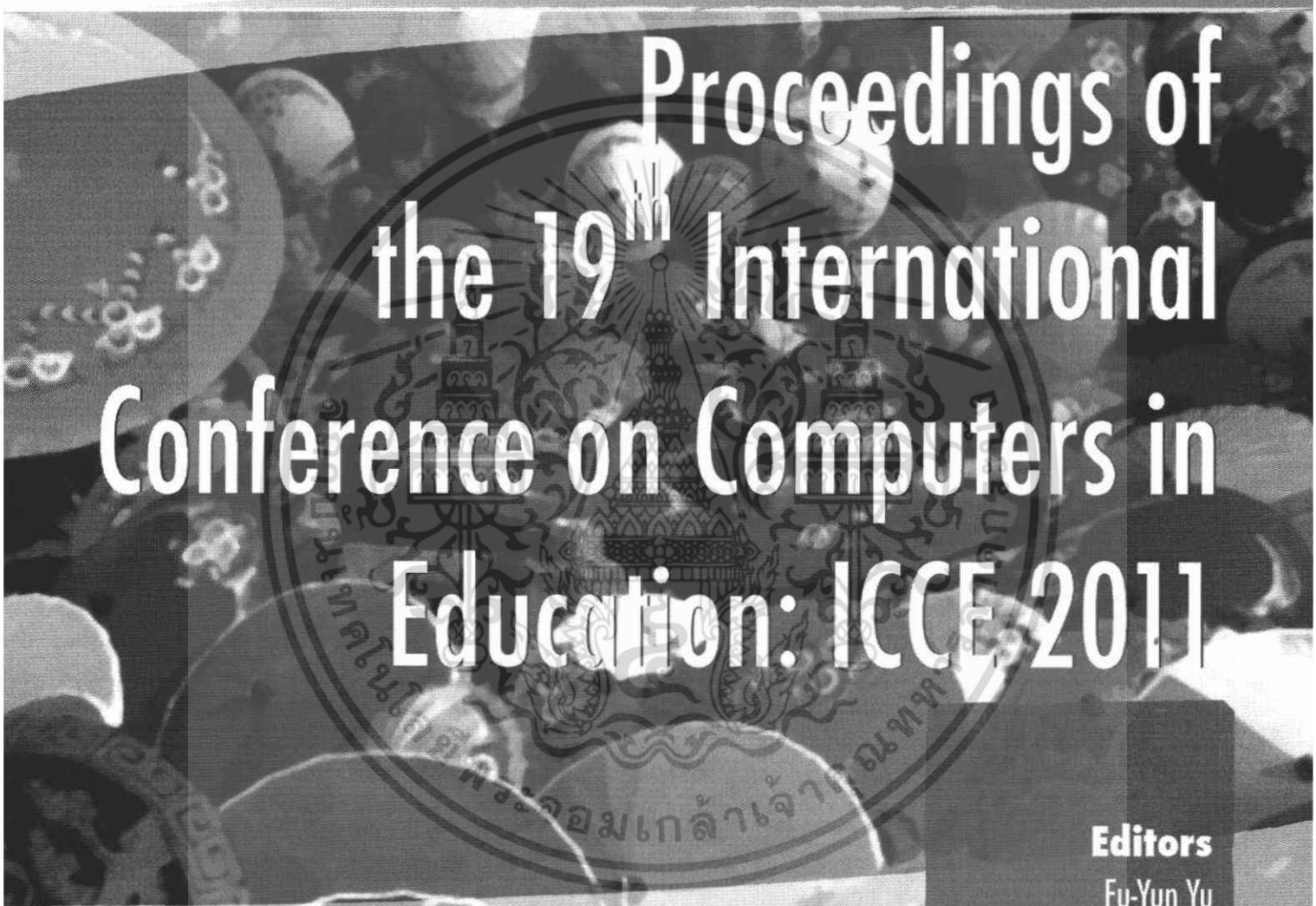
[10] S. Haykin, 2008. "Neural Networks and Learning Machines ". Prentice Hall, USA.

[11] R. Kohavi, 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143.

ICCE

The 19th International Conference on Computers in Education 2011

28 November - 2 December • Chiang Mai, Thailand



Proceedings of the 19th International Conference on Computers in Education: ICCE 2011

Editors

Fu-Yun Yu

Tsukasa Hirashima

Thepchai Supnithi

Gautam Biswas

APSCE

NECTEC a member of NSTDA

TCU Thailand Cyber University

MAEJO UNIVERSITY

TABLE OF CONTENTS

T. Hirashima et al. (Eds.) (2011). Proceedings of the 19th International Conference on Computers in Education. Chiang Mai, Thailand: Asia-Pacific Society for Computers in Education

Statistical Level Checker with Personalised English Passage Suggestion

Wasan NA CHAI[ab] **Taneth RUANGRAJITPAKORN[b]**

Nualsawat HIRANSAKOLWONG[a] **Thepchai SUPNITHI[b]**

[a]Mathematics and Computer Science department,
King Mongkut's Institute of Technology Ladkrabang, Thailand
[b]Human Language Technology Laboratory,
National Electronic and Computer Technology Center, Thailand
wasannachai@gmail.com, khnualsa@kmitl.ac.th
{taneth.ruangrajitpakorn, thepchai.supnithi}@nectec.or.th

Abstract: In this paper, a system to classify a readability level of English reading passage and to match student personal interest is purposed. Student model is applied to collect student information for selecting their preferable passage topic. Statistical passage level checker is implemented to match student readability level with passage difficulty by using neural network. Three linguistic features, syllable, vocabulary and sentence complexity, are chosen to distinguish a difficulty difference among passage level. The best accuracy gained by the system is 86.25% and the constantly reliable feature for this task is a sentence complexity of the passage.

Keywords: Readability level checker, English reading passage, English learning, personalisation, neural network

1. Introduction

English class is one of the most boredom subject for Thai students since Thai children are not familiar with inflection, syntactic word order, and grammar learning therefore they become idle and inactive in class. Furthermore, their reading passages become tiresome since each student has individually preferable topic and they tend to lose their learning motivation to read an assigned non-interested passage. Letting students choose reading-passage by themselves also leads to significant burden for instructors to scope an appropriate level of those passages.

In language learning, readability checker tool or passage grading system are one of the important application that assists students and instructors in terms of reducing a load to select a proper readability level of reading passages. The major issue is that most of them was implemented on rule-based approach and the designed rules are reckoned for an English native speaker. The rule-based systems and methods including Flesch reading easy formula [1], Kincaid formula [2], SMOG-grading [3] and Fox index [4] are rigid and can hardly be applied to students who study an English language as their foreigner language since the level of English understanding and skill are rather different based on each country standard of language learning.

Recently, the framework for passage grading system using statistical approach was purposed [5]. However, the mentioned system was a framework which exploits three

linguistic features; syllable, vocabulary and sentence complexity score, along with a conditional random field (CRF) as their machine learning to create a model of the passage level. Although CRF is reliable one among other machine learning techniques for its discriminative training, it has a specification to number management since CRF recognises an input as string, not an integer. So far, an experiment result of the framework has not been published. Another study on a passage grading system using supervised learning by a neural network [6] was later reported. It applied the same linguistic features as the above mentioned framework but the machine learning was altered to a neural network. They applied a neural network to generate three models based on each linguistic feature and exploited those models separately to level a reading passage based on an academic level. The limitation of the system is not much sufficient accuracy as around 80%. The statistical systems work properly in practical but the load falls to students who have to search a reading passage by themselves and they occasionally conduct a searching again if the system returns an unsuitable readability level to them.

The question to be solved in this paper is to find a solution for matching student readability with their preferable topic. Furthermore, to improve an accuracy of passage level checker, we extend the existing statistical passage level checker system using neural-network by integrating the three features into a single model and compare the result between those two methods. Last, each feature is focused to compare the efficiency and reliability among them.

2. Statistical Level Checker with Personalised English Passage Suggestion

Statistical Level Checker with Personalised English Passage Suggestion is an automatic system for matching student's readability and a difficulty level of an English reading passage with student personalisation. The system consists of two main parts; student model and passage level checker. Student model represents student information for selecting interesting passage and improving English skill for individual student while passage level checker is a tool to examine a compatible difficulty of reading passage to student. An overview of the tool is sketched in Figure 1.

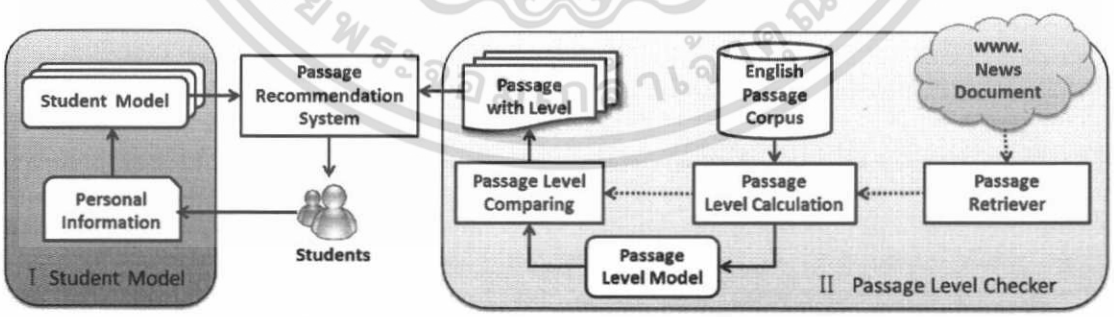


Figure 1. An overview of the system

2.1 Student Model

Students model is a representation of student personal information in several aspects. To recognise students' interest and performance in English learning, student model is designed to consist of three parts; 1) profile information, 2) interest information and 3) competency information.

The profile information is a personal profile that collects details about student name, gender, age and current academic level. The interest information gathers student interested topics and activities including their desired occupation, hobby, sport, favourite food and beverage, favourite song genre and movie, etc. The competency information is a collection of student competence from test results. Competency tests are designed to seek out student strong and weak fluency of English reading skills. The result will help on selecting the passage full of their weak linguistic feature which will improve such skill to individual. An example of student model is shown in Table 1. With the information, a passage retriever and a similar topic recommendation has been applied to select an appropriate passage to satisfy student interested topic.

Table 1. Examples of four sentence types with underlined criteria

Type	Subject	Student A			Student B		
profile	name	Peera Chareounsap			Chutima Lakprae		
	gender	male			female		
	age	15			14		
	academic level	level 9			level 8		
interest	desired occupation	veterinarian, astronaut			dentist, scientist		
	hobby	collecting stamp			drawing		
	sport	football, tennis			badminton, swimming		
	favourite food	pizza, noodle			noodle, ice-cream		
	favourite beverage	soda			fruit juice		
	favourite music genre	rock, pop			pop, r&b		
	favourite movie	thrill, sci-fi			romantic, comedy		
competency	test result	time	vocab	grammar	summary	vocab	grammar	summary
		1st	17/30	12/30	18/30	17/30	12/30	18/30
		2nd	19/30	15/30	17/30	19/30	15/30	17/30
		3rd	20/30	14/30	20/30	20/30	14/30	20/30

2.2 Passage Level Checker

Passage level checker is a tool that automatically identifies a readability level of an English reading passage by comparing to a passage corpus. Three linguistic features; vocabulary, syllable and sentence complexity, are exploited to distinguish the differences among passage-level. In training process, level models based on the number of levels from a passage corpus is generated. The models are afterwards used to determine a level of a target passage and the tool returns its grade level as a result. A diagram of the passage level checker is demonstrated in Figure 2.

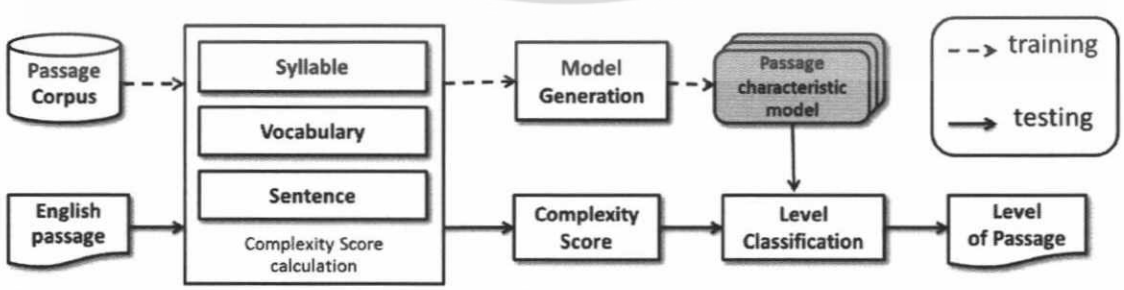


Figure 2. An overview of the passage level checker tool

In the pre-process for constructing an English reading passage corpus, Englsih passages are word-segmented by white space and sentence are divided by full-stop,

question mark, and exclamation mark punctuation. An abbreviated form of auxiliary verb is expanded fully, for instance, "*he'll*" is transformed to "*he will*".

2.2.1 Syllable Complexity Score Calculation

A syllable complexity is a quality of average difficulty of words existing in the reading passage. AJAX syllable counter [7] is applied to count a syllable amount of each word. An average number of syllable for each passage is calculated by (1)

$$F1 = \frac{\sum_{i=1}^{n}(n_{syl_i})}{W} \qquad (1)$$

where n_{syl_i} is the number of syllable of word *i*th and W is the total number of words in a passage.

2.2.2 Vocabulary Complexity Score Calculation

A vocabulary complexity is a measurement of a lexical meaning difficulty in the passage context. To calculate vocabulary complexity, word classes which are content and function word are separately concerned because of their different significance.

A content word shows a stable lexical meaning and it is an open-class word which opens to possibilities for expansion. On the other hand, a function word is a word that contains little lexical meaning, but instead serves to express grammatical relationships with other words and function words are relatively small number of items. Moreover, content words are variable in form due to inflection. Therefore, words in a passage are split into two classes and handled separately in word level examination. For content words, lemmas[1] are extracted by Morpha [8][9], a lemmatisation tool, to prevent non-matching inflected forms. To some degree, content words are recognised to be more difficult than function words thus a parameter for a function word is set to 1.0 while a parameter for a content word is set greater to 1.5.

Beside of named-entities, unknown words that do not match the reference level word list are treated as the highest level since they are inclined to be a domain-specific word or specialised technical term. A process of vocabulary complexity score calculation is sketched in Figure 3.

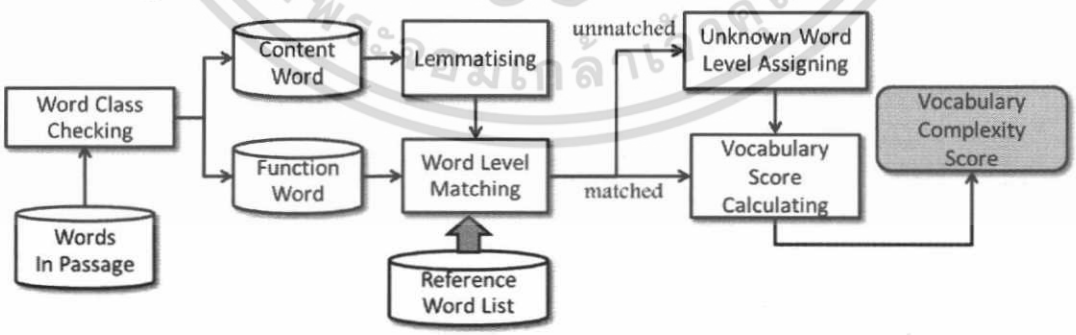


Figure 3. A process of vocabulary complexity score calculation

With above mentioned criteria, a vocabulary complexity score is computed by (2).

1 A lemma refers to the canonical form, dictionary form, or citation form of a word, e.g., in English, die, dies, died, and dying are forms of the same lexeme, with "die" as their lemma. It is different from a word stem which all affixes are removed.

$$F2=\frac{\sum_{i=1}^{n}[(Lv_{c_i}\cdot 1\cdot n_c)+(Lv_{f_i}\cdot 1.5\cdot n_f)]}{\sum_{i=1}^{n}(Lv_{w_i}\cdot W_i)} \qquad (2)$$

where Lv refers to a level of a word in reference list, c indicates a content word, f is a function word, n is a frequency, w refers to any kind of word and W_i is a frequency of i^{th} word.

2.2.3 Sentence Complexity Score Calculation

Sentence complexity is a difficulty of sentences in a passage. Apparently, the more clauses a sentence has, the more difficulty it gains. Basically, a sentence type in English falls into four types which are simple sentence (S), compound sentence (CP), complex sentence (CX) and compound-complex sentence (CPX). The main criteria used in this process is to capture an existence and a type of conjunction and clause marker within the sentence with co-occurring punctuation(s). The complexity score of sentence type is computed based on an amount of complexness of clause(s) by using (3).

$$F3=\frac{\sum_{i=1}^{n}(S\cdot CP^{N_{CP}}\cdot CX^{N_{CX}})}{Total_{sentence}} \qquad (3)$$

where S refers to a simple sentence, CX indicates a complex sentence and CP is a compound sentence. N_{CX} is a number of a literation of a complex sentence and N_{CP} is a literation number of compound sentence. In case of CPX, it is counted if both CP and CX exist at least one. The examples of sentence types within the corpus are shown in Table 2.

Table 2. Examples of four sentence types with underlined criteria

Type	Examples
S	- The boy bought that books from the store. - Coral provides good hiding places for fish.
CP	- He sees the recycling truck, and he also sees Janey behind it. - It is really dark in the art gallery, but Harry has a light
CX	- Dinosaurs lived on Earth long before there were any people. - When he pushed the last sign into place, the seven lines of light shone into the middle where Henry used to stand on.
CPX	- The man wants the painting, but Harry doesn't have it since the painting has been stolen. - The two friends silently agreed, but their faces showed no fear though the beast ran toward them.

2.2.4 Model Generation and Level Classification

In the former implementation of statistical passage grading system [6], Neural-network [10] is applied to generate three models to determine the level of a passage. Currently, we alter the model generation in two steps by creating a model of features and apply the obtained model into vector to generate a level model by neural-network again. Figure 4 shows a comparison between former model generation method and the new method.

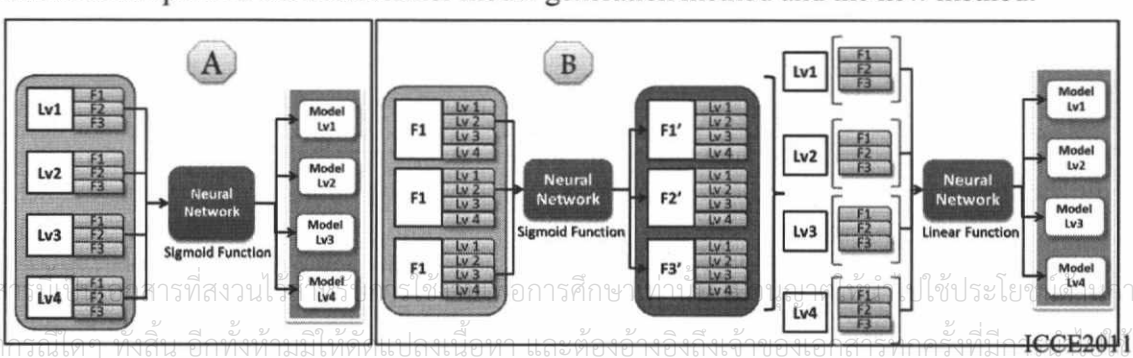


Figure 4. Flows between former model generation and the vector model generation

The difference of Figure 4A and Figure 4B is that the proposed vector method in figure 4B could explicitly return an available method to examine feature impact for each level. Once a passage model is obtained, it is used to classify a level of the target reading passage by the use of neural-network to determine probability as shown in Figure 5.

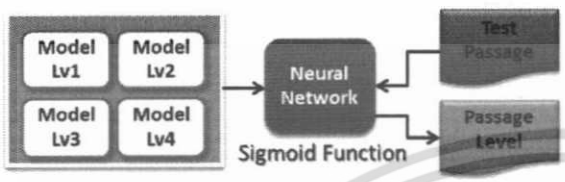


Figure 5. A level classification process

2.3 Integration of Student Model and Passage Level Checker

Once student model and result of passage level checker are gained, the recommendation seeks keywords and a domain topic from the retrieved passage and matches the found information with the interested item of individual student. The final result of the system is an English reading passage with preferable topic and suitable readability level.

There are two beneficial aspects for incorporating student model and passage level checker. The former is to help students to enjoy their reading with their interested topic passage. The latter is to select a reading passage with an appropriate level that suits student readability. Moreover, the system has an optional function to choose a passage which is compatible to student weak skills by observing the competency information to especially improve such skill. This will improve Thai student in English learning with their better attention and motivation and it also becomes a student oriented learning.

3. Passage Level Checker Experiment

3.1 Passage Corpus and Reference Word List

Reading passages used for training and testing were collected from reading passages and supplementary reading passage assigned in school class approved by the Ministry of Education of Thailand. The chosen passages are for Thai students who learn English as a foreigner language. The reading passages are divided into four grade-levels based on an academic grade of Thailand; junior primary school (grade 1-3), primary school (grade 4-6), junior high school (grade 7-9) and high school (grade 10-12). The number of passages for each level is 200 reading passages and the total number is 800.

To examine word level, reference word list is collected from vocabulary list approved by the Ministry of Education of Thailand. Table 3 shows a statistic of the reference word list of each grade-level.

Table 3. A statistic of reference word list from the Ministry of Education of Thailand

grade level	content word	function word	total
junior primary school	560	33	593
primary school	2,111	71	2,182
junior high school	3,566	82	3,648
high school	3,802	57	3,859
total	10,039	243	10,282

3.2 Experiment Setting and Result

To estimate an accuracy, 5-fold cross-validation is applied. Two methods of model generation; three model method and single metric model method, are tested separately. The comparison result between former model generation method and our method is shown in Table 4. To compare efficiency of single feature, accurate result of each feature and combination of features separated by level are shown in Table 5 where *Sy*, *Vo* and *Se* stand for syllable complexity, vocabulary complexity and sentence complexity respectively. A total amount and percentage of accuracy gained from each feature focused only from the correct results are given in Table 6.

Table 4. A result between former model generation method and the purposed method

	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg.
Former method	86.72%	84.38%	83.59%	80.47%	82.03%	83.44%
Metric method	83.59%	87.50%	88.28%	82.81%	85.78%	**86.25%**

Table 5. An accurate result of each feature and combination of features

		Sy	Vo	Se	Sy+Vo	Sy+Se	Vo+SE	All	Sum
Amount of Correct passage	Level 1	2	6	4	23	20	37	47	139
	Level 2	2	6	2	20	22	41	46	139
	Level 3	2	3	6	16	38	36	36	137
	Level 4	1	1	9	15	36	35	37	134
	Sum	7	16	21	74	116	149	166	549

Table 6. A total amount and percentage of accuracy gained from each feature

	Syllable (amount|%)		Vocabulary (amount|%)		Sentence (amount|%)	
Level 1	92	66.19%	113	81.29%	108	77.70%
Level 2	90	64.75%	113	81.29%	111	79.86%
Level 3	92	67.15%	91	66.42%	116	84.67%
Level 4	89	66.42%	88	65.67%	117	87.31%
Sum	363	66.12%	405	73.77%	452	82.33%

4. Discussion

The proposed model generation method does not give sufficiently higher accuracy than the former method. However, it allows us to directly investigate the impact of each feature rather than the former one which is hard to access to feature tuning.

From comparison of each feature, the sentence complexity constantly shows reliability for classifying passage level and the syllable complexity is a moderate feature in this task. For the vocabulary complexity, its accuracy obviously depends on a level of a passage. The vocabulary complexity performs greatly for lower levels while sentence complexity shows decent potential on higher level. Since the lower level passages (level 1 and 2) contain reoccurred simple vocabularies in the easy conversation style and the number of lexicons is small, vocabulary complexity can capture them simply and returns the most accurate result. In the other hand, the higher level passages (level 3 and 4) contain several lexical meaning words and the variety of them based on different domain cause the performance of vocabulary complexity to certainly drop. Moreover from error observation, we found two major issues which are unknown word issue and multi-sense

word issue (polysemy) to emphatically lessen accuracy of vocabulary complexity. The former issue is caused by the missing word from reference word list. Many simple and general words are absent from the list especially a noun, for instance, "dragonfly", "coral", "glove", "dinosaur", "motive", "helmet", etc. These unknown words are ranked to highest level and cause the system to determine a passage containing them to higher level than its realistic level. Since the reference word list from the Ministry of Education of Thailand is not reliable because of non-coverage issue, the solution will fall to garner the words from the passage corpus itself and rank them by existence frequency in each level. The latter issue is a word with multiple meanings. They cannot be handled in the system efficiently since the system recognised them as they are the same and treat them as its lowest level in the reference word list. This causes the vocabulary complexity score calculation to give a lower level to a passage than it should be.

For syllable complexity, the length of syllable is not a certain measurement for reading difficulty since some short word can be more difficult than the longer syllable word, for instance, the word "woe" which possesses one syllable is definitely harder to understand for English learners than the word "butterfly" which counts as three syllables. Therefore, the performance of the feature is not much reliable for passage level determining task.

5. Conclusion and Future Work

In this paper, we purpose a statistical passage level classification system which provide an English reading passage with proper readability level and preferable topic for Thai student. Student model is applied to detect student interested topic of reading and their readability whilst passage level checker provides a level approval to filter a reading passage that does not suit student readability. The statistical passage level checker distinguishes a readability difficulty from different level by calculating three linguistic features which are syllable complexity, vocabulary complexity and sentence complexity. Neural-network is exploited to generate a level characteristic model based on three above-mentioned features from a reading passage corpus to prevent a rigidity of inflexible criteria for different English learning standard. From the experiment result, the average accuracy is 86.25% while the sentence complexity score shows a potential on passage level determination for a single feature. The variable accuracy depending on passage level falls to vocabulary complexity score which encounters a matching issue from reference word list in terms of polysemous ambiguity and non-coverage lexical entry. To improve the passage level checker, we plan to add word sense disambiguation to solve polysemous word issue. A better method to gather reference word list will be researched for better vocabulary complexity score calculation. Furthermore, new linguistic features such as speech type (direct speech and indirect speech) and idiom usage will be attached to specify more accurate readability difficulty. For recommending interested passage, a topic selection will concern with more implicit personal information, such as parent marriage status or their relationship with community, to prevent suggesting a non-suitable passage.

References

[1] Flesch, Rudolf . "A New Readability Yardstick," Journal of Applied Psychology, 32, 1948, 221-233.

[2] Luo Si and Jamie Callan.2001. A Statistical Model for Scientific Readability. In Proceedings of the tenth international conference on Information and knowledge management.

[3] McLaughlin, G. Harry (May 1969). "SMOG Grading — a New Readability Formula". Journal of Reading 12 (8): 639–646.

[4] Stephen Fuller, PharmD, Cheryl Horlen, PharmD, Robert Cisneros, PhD, and Tonja Merz, PharmD. 2007. "Pharmacy Students' Reading Ability and the Readability of Required Reading Materials". American Journal of Pharmaceutical Education 2007; 71 (6) Article 111.
[5] Wasan Na Chai, Taneth Ruangrajitpakorn, Thepchai Supnithi. 2010. A Statistical Approach on Automatic Passage Level Checking Framework for English Learner. In Proceedings of ICCE 2010.
[6] Wasan Na Chai, Naulsawat Hiransakolwong, Taneth Ruangrajitpakorn, Thepchai Supnithi. 2010. Passage Grading System Using Supervised Learning. In Proceedings of KICSS 2010.
[7] AJAX Syllable Counter : available at http://www.russellmcveigh.info/content/html/syllablecounter.php
[8] G. Minnen, J. Carroll and D. Pearce. (2000). 'Robust, applied morphological generation'. In Proceedings of the 1st International Natural Language Generation Conference, Mitzpe Ramon, Israel. 201-208.
[9] Morphological and Orthographic Tools for English (morpha): available at http://www.informatics.sussex.ac.uk/research/groups/nlp/carroll/morph.html
[10] Hornik, K., Stinchcombe, M., White, H., 1989, "Multilayer feedforward networks are Universal Approximators," Neural Networks 2: 359-366.

ประวัติผู้เขียน

ชื่อ – สกุล	นายวสันต์ ณ ชัย
วัน เดือน ปีเกิด	11 กรกฎาคม 2524
ที่อยู่	11/53 ถ.เอกาทศรฐ อ.เมือง จ.พิษณุโลก 65000
ประวัติการศึกษา	
2549	จบการศึกษาวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร
2550- ปัจจุบัน	กำลังศึกษาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง