บทที่ 5

การออกแบบกระบวนการขั้นตอนของงานวิจัย

1. ความเบื้องต้น

การระบุคำหรือวลีสำคัญให้ได้ความแม่นยำและถูกต้องมีประสิทธิภาพนั้น จะขึ้นอยู่กับเทคนิคในการตัดแยกและสกัดคำในเอกสารภาษาไทย กฎการให้ค่าน้ำหนักคำหรือวลีสำคัญที่ใช้ในการสอนโครงข่าย หรือที่เรียกว่า กฎการเรียนรู้ รวมถึงโมเดลที่ได้จากการเรียนรู้ด้วยข้อมูลที่เหมาะสมและมีประสิทธิภาพในระดับที่น่าพอใจ

โดยหลักการของการเกาะเกี่ยวใจความสำคัญในเอกสารภาษาไทยนั้น ใจความสำคัญของแต่ละเอกสารอาจอยู่ในตำแหน่งที่แตกต่างกันออกไป โดยการอ่านเอกสารที่กระทำโดยมนุษย์หรือโดยบุคคลที่เชี่ยวชาญในสาขาวิชานั้น หลังจากการอ่านเอกสารแล้วนั้นจะสามารถทำการระบุถึงตำแหน่งที่อยู่ของใจความสำคัญ และสามารถบอกได้ว่าส่วนใดหรือคำวลีใดเป็นใจความสำคัญของเอกสาร แต่ระบบการค้นหาเป็นระบบคอมพิวเตอร์หรือเครื่องจักรที่ไม่สามารถวิเคราะห์ได้เหมือนกับมนุษย์ ดังนั้นถ้าจะให้ระบบสามารถแยกแยะหรือค้นหาคำหรือวลีสำคัญในเอกสารได้นั้น ควรจะต้องมีการสอนให้ระบบได้เรียนรู้ก่อนว่าส่วนใด(ตำแหน่งใด)คือคำหรือวลีที่มีค่าน้ำหนักที่เหมาะสมจะเป็นตัวแทนสำคัญของเอกสาร โดยอาศัยโครงข่ายประสาทเทียมในการสกัดคำหรือวลีสำคัญของเอกสารภาษาไทย

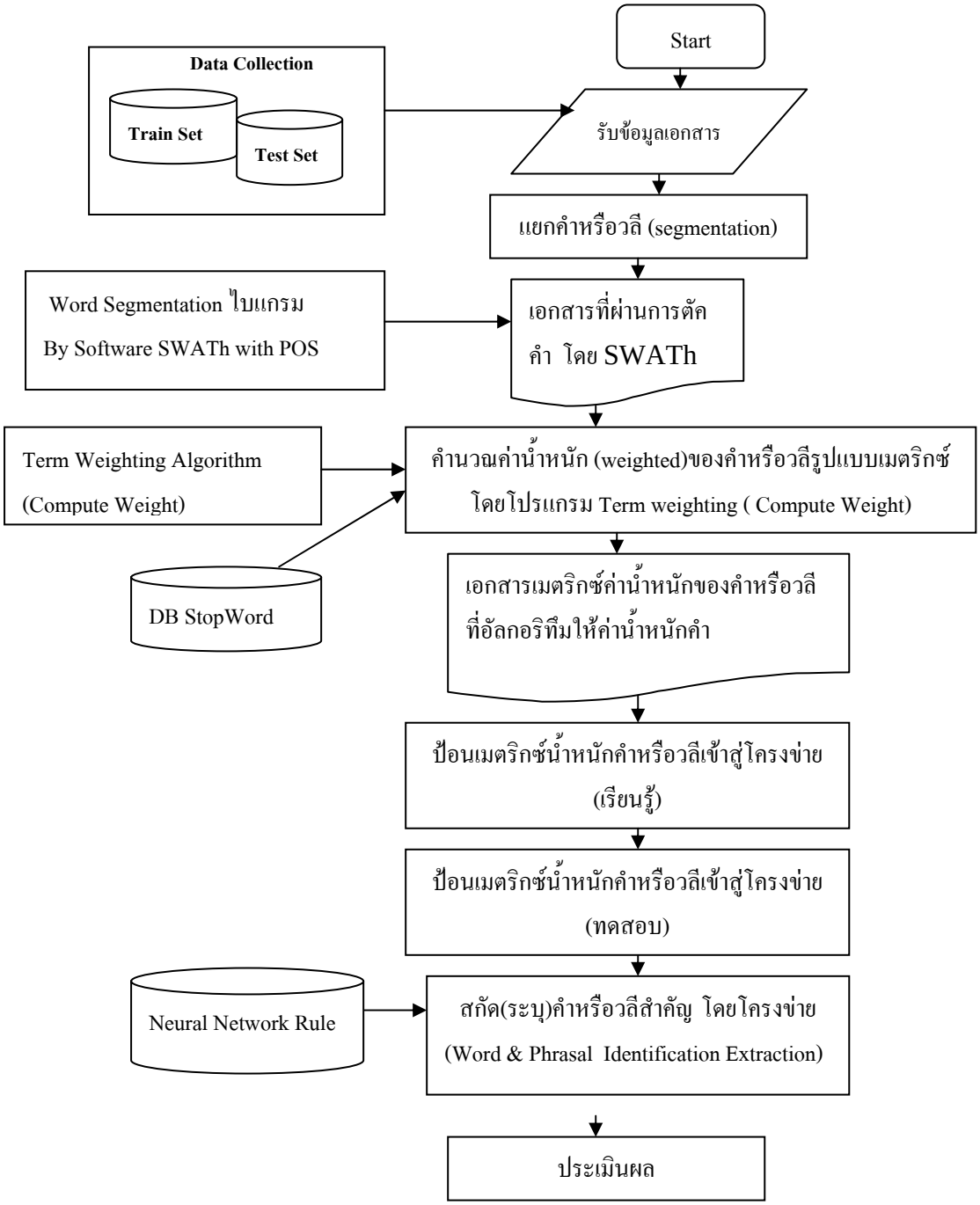
2. แนวคิดงานวิจัยการสกัดวลีสำคัญในเอกสารภาษาไทย

ในงานวิจัยนี้ ผู้วิจัยมุ่งที่จะศึกษาหลักการของไวยากรณ์ภาษาไทยเพื่อนำมาปรับปรุงประสิทธิภาพของระบบการค้นคืนเอกสารที่เป็นภาษาไทย โดยจะนำเสนออัลกอริทึมและกฎการเรียนรู้ที่อ้างหลักไวยากรณ์ภาษาไทย สำหรับการค้นคืนสารสนเทศที่มีความถูกต้องและแม่นยำมากขึ้นโดยผ่านโครงข่ายประสาทเทียม ในงานวิจัยนี้กำหนดวิธีการดำเนินงานไว้ดังนี้

1. การออกแบบรูปแบบของข้อมูล
2. การจัดเก็บและจัดเตรียมข้อมูล

3. ขั้นตอนการตัดคำด้วยหลักการแบบ ไบแกรม ที่อ้างอิงหลักไวยากรณ์ภาษาไทย
4. ขั้นตอนการให้ค่าน้ำหนักคำจากอัลกอริทึมกฎการเรียนรู้ของหลักทางไวยากรณ์
5. การระบุคำหรือวลีสำคัญในเอกสารภาษาไทยโดยโครงข่ายประสาทเทียม
6. การประเมินผลสำคัญ

3. หลักการทำงาน



ภาพประกอบที่ 5-1 แสดงโครงสร้างของการทำงานของการระบุคำหรือวลีสำคัญ

4. กระบวนการทำงานของงานวิจัย

ในการออกแบบกระบวนการทำงานของงานวิจัย มีขั้นตอนของการทำงาน คือ ทำการเรียนรู้ข้อมูลและทำการสกัดหรือระบุคำหรือวลีสำคัญในเอกสารนั้น ดังภาพประกอบ 5-1 ซึ่งจะแสดงให้เห็นลักษณะงานโดยรวมของการระบุคำหรือวลีสำคัญในเอกสารภาษาไทย โดยการทำงานต้องอาศัยขั้นตอนที่สำคัญมาประกอบกัน โดยมีขั้นตอนในการทำงานคล้ายกัน คือ การรับข้อมูลคำหรือวลี การตัดแยกคำ การให้ค่าน้ำหนักของคำ การใช้โครงข่ายประสาทเทียมเรียนรู้คำหรือวลี และการใช้โครงข่ายประสาทเทียมในการระบุคำหรือวลีสำคัญในเอกสาร

4.1 การเตรียมข้อมูล (Data Pre-processing)

ฐานข้อมูลสำหรับเป็นข้อมูลฝึกสอนและข้อมูลทดสอบนั้น ผู้วิจัยเลือกนำมาจากคลังข้อความที่นำมาใช้งานวิจัยนั้นได้นำมาจากคลังข้อมูลออร์คิด โดยลักษณะของบทความที่นำมาใช้สร้างคลังข้อความนั้น ได้นำมาจากรายงานการประชุมวิชาการของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

ผู้วิจัยรวบรวมข้อความภาษาไทยจากคลังข้อมูลออร์คิด เพื่อนำมาใช้เป็นคลังข้อมูลในวิทยานิพนธ์ฉบับนี้ โดยจะแบ่งเป็น 2 ส่วน คือ

- คลังข้อมูลฝึกสอน (training corpus) ซึ่งเป็นข้อความที่ยังไม่ได้ตัดคำและกำกับหมวดคำ เพื่อสำหรับให้โปรแกรมเรียนรู้
- และคลังข้อมูลทดสอบ (test corpus) โดยเป็นข้อความที่ยังไม่ได้ทำการตัดคำและกำกับหมวดคำไว้

โดยมีขั้นตอนในการเตรียมข้อมูลสำหรับฝึกสอนและทดสอบ ดังนี้

4.1.1 ขั้นตอนการแปลงข้อมูล

ในขั้นตอนนี้ ผู้วิจัยได้ทำการถ่ายโอนข้อมูลและแปลงข้อมูลเอกสารให้อยู่ในรูปแฟ้มข้อมูลในลักษณะที่เป็นข้อความจากข้อมูลเดิมในคลังข้อมูลออร์คิด โดยตัวอย่างของแฟ้มข้อมูลเดิมในคลังข้อมูลออร์คิด แสดงในรูปที่ 5-2 โดยการรับข้อมูลเข้าสู่ระบบ ของโปรแกรมตัดแยกคำนั้นสามารถรองรับเอกสารที่จะตัดคำได้ 2 รูปแบบ คือ เอกสาร .txt และ เอกสาร.doc

ผู้วิจัยจึงเลือกพิจารณาที่จะแปลงข้อมูลเอกสารทดสอบนั้นให้อยู่ในรูปของเอกสาร .txt ดังภาพ 5-3 สำหรับนำเข้าสู่กระบวนการตัดแยกคำ ดังที่จะกล่าวในหัวข้อต่อไป

```
(1)
%TTitle: การประชุมทางวิชาการ ครั้งที่ 1
%ETitle: [1st Annual Conference]
%TAuthor:
%EAuthor:
%TInbook: การประชุมทางวิชาการ ครั้งที่ 1, โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์, ปีงบประมาณ 2531, เล่ม 1
%EInbook: The 1st Annual Conference, Electronics and Computer Research and Development Project
%TPublisher: ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, กระทรวงวิทยาศาสตร์ เทคโนโลยีและการพลังงาน
%EPublisher: National Electronics and Computer Technology Center, Ministry of Science, Technology
%Page:
%Year: 1989
%File:
#P1
#1
การประชุมทางวิชาการ ครั้งที่ 1//
การ/FIXN
ประชุม/VACT
ทาง/NCMN
วิชาการ/NCMN
<space>/PUNC
ครั้ง/CFQC
ที่ 1/DONM
//
#2
โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์//
โครงการวิจัยและพัฒนา/NCMN
อิเล็กทรอนิกส์/NCMN
และ/JCRG
คอมพิวเตอร์/NCMN
```

ภาพประกอบที่ 5-2 ตัวอย่างประโยคภายในคลังข้อความออร์คิด

TTitle: ระบบเตรียมข้อมูลคอมพิวเตอร์อัตโนมัติภาษาไทยและภาษาอังกฤษ

ETitle: Automatic Recognition of Thai - English Characters System

TAuthor: ดร.ชม กิ้มปาน วัชระ ฉัตรวิริยะและ สุรสิทธิ์ ราตรี

TInbook: การประชุมทางวิชาการ ครั้งที่ 1, โครงการวิจัยและพัฒนาอิเล็กทรอนิกส์และคอมพิวเตอร์, ปีงบประมาณ 2531, เล่ม 1

ABSTRACT

บทความผลงานวิจัยนี้ เสนอระบบเตรียมข้อมูลคอมพิวเตอร์อัตโนมัติภาษาไทยและภาษาอังกฤษ (Automatic Recognition of Thai-English Characters System) ข้อมูลที่พิมพ์หรือเขียนอยู่บนกระดาษธรรมดาที่เป็นตัวอักษรภาษาไทยหรือตัวภาษาอังกฤษ จะถูกป้อนผ่านเครื่อง Image Scanner ลักษณะลายเส้นของตัวอักษรจะถูกแทนด้วยรหัส 1 ลักษณะพื้นเบื้องหลังของลายเส้นจะถูกแทนด้วยรหัส 0 ข้อมูลที่ได้จะถูกจัดเก็บไว้ในลักษณะ Image file จากนั้นกระบวนการ Segmentation เป็นวิธีการดึงข้อมูลที่เป็นลักษณะ (pattern) ของตัวอักษรแต่ละตัวออกจาก Image file ซึ่งในการวิจัยนี้ใช้วิธีการติดตามรอยขอบมาใช้ ลักษณะข้อมูลของแต่ละตัวอักษรจะถูกจัดเก็บเป็นไฟล์ๆ ไป จากนั้นก็จะถึงกระบวนการรู้จำ (Recognition) เป็นการนำข้อมูลในแต่ละไฟล์ที่เป็นข้อมูลของแต่ละตัวอักษรมาทำการวิเคราะห์ว่าข้อมูลนั้นเป็นตัวอักษรอะไร วิธีการที่ใช้ในงานวิจัยนี้ เลือกใช้อยู่สองวิธี คือ Feature concentation method และ Recognition by using K-L expansion method ผลของการวิเคราะห์จะใช้รหัส ASCII สำหรับแต่ละตัวอักษร แล้วจัดเก็บเป็น Text file สำหรับนำไปใช้งานต่อไป ซอฟต์แวร์ของกระบวนการ Segmentation และกระบวนการ Recognition จะอยู่ในลักษณะของ Machine code บน Card พิเศษที่สร้างขึ้นมา สำหรับนำไปใช้ร่วมกับไมโครคอมพิวเตอร์แบบให้

ภาพประกอบที่ 5-3 ตัวอย่างข้อมูลหลังการแปลงจากคลังข้อมูล เพื่อเตรียมตัดคำ

4.2 การเลือกใช้โปรแกรตัดแยกคำ ที่อ้างอิงกับหลักไวยากรณ์ภาษาไทย

ในขั้นตอนนี้ เป็นการกล่าวถึงรายละเอียดของการตัดแยกคำหรือวลีในเอกสารฝึกสอนและเอกสารทดสอบ เพราะลักษณะของการเขียนในเอกสารภาษาไทยนั้นจะเขียนคำในแต่ละประโยคยาวติดกันไปจนกว่าจะจบประโยค จึงทำให้ไม่สามารถทราบได้เลยว่าส่วนใดคือตำแหน่งสิ้นสุดของคำหรือวลี ด้วยเหตุนี้ในงานด้านประมวลผลภาษาธรรมชาตินั้น จึงจำเป็นต้องทำการตัดแยกคำหรือวลีออกจากกันก่อนในเบื้องต้น โดยผู้วิจัยจะทำการตัดคำและกำกับหมวดคำให้กับแต่ละคำหรือวลีในเอกสารทดสอบตามเกณฑ์การตัดคำและชุดหมวดคำของโปรแกรมตัดแยกคำ SWATh (ไพศาล ,2541)

4.2.1 เทคนิคการตัดคำไทยที่ผ่านมา

การตัดคำในปัจจุบันนั้น มีด้วยกัน 3 เทคนิค คือการตัดคำแบบใช้กฎ , การตัดคำแบบใช้คลังข้อมูล และการตัดคำแบบอาศัยพจนานุกรม

4.2.1.1 หลักการตัดคำแบบใช้กฎ

หลักการตัดคำแบบใช้กฎ นั้น เป็นการพัฒนาระบบการตัดคำภาษาไทยโดยใช้วิธีการตรวจสอบกฎเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร การเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการบ่งชี้ขอบของคำ

- ข้อด้อยของเทคนิคการตัดคำแบบอาศัยกฎ

โดยถ้ากฎมีจำนวนน้อยเกินก็ทำให้ตัดแยกคำได้ไม่ดีเพราะผลของการตัดคำอาจได้เป็นกลุ่มคำที่สามารถตัดคำแยกย่อยออกไปได้อีก หรือถ้ากฎมีจำนวนมากเกินไปก็อาจทำให้ใช้เวลาในการประมวลผลนาน และการตัดคำแบบใช้กฎนี้ ผลค่าความถูกต้องในการตัดแยกคำมีค่าความถูกต้องค่อนข้างต่ำ ดังจะเห็นได้จากงานวิจัยของยุพิน ไทยรัตนานนท์ (1981) เป็นงานวิจัยการตัดพยางค์ โดยการใช้กฎในการตัดพยางค์ ซึ่งกฎต่าง ๆ ที่สร้างขึ้นมานั้นโดยอาศัยหลักไวยากรณ์ภาษาไทย แต่ก็จะมีปัญหาในการสร้างกฎเพราะมีบางพยางค์ไม่เป็นไปตามกฎที่ตั้งไว้ และกฎต่าง ๆ ที่สร้างขึ้นมานั้น ยากต่อการเพิ่มหรือแก้ไข โดยจากการทดสอบปรากฏว่าผลการตัดคำให้ค่าความถูกต้อง 85.00 % โดยจะเห็นได้ว่างานวิจัยที่ใช้เทคนิคการตัดคำด้วยกฎนั้น ยังได้ค่าความถูกต้องที่น้อยกว่าเทคนิคการตัดคำแบบอื่น ๆ ที่จะกล่าวต่อไป ผู้วิจัยจึงไม่เลือกเทคนิคการตัดคำเทคนิคนี้

4.2.1.2 หลักการตัดคำแบบใช้คลังข้อมูล

หลักการตัดคำแบบแบบใช้คลังข้อมูล นั้น คือ การพัฒนาคลังข้อความขนาดใหญ่ที่จัดเก็บความรู้ต่าง ๆ อาทิเช่น สถิติการใช้คำภายในคลังข้อความและลักษณะไวยากรณ์ช่วยในการตัดแยกคำ ทำให้มีความถูกต้องในการตัดแยกคำมากยิ่งขึ้น โดยเทคนิคการตัดคำแบบอาศัยคลังข้อมูลนี้เทคนิคหนึ่งที่ช่วยในการแก้ไขปัญหาความกำกวมในการตัดแยกคำ และทำให้มีความถูกต้องมากยิ่งขึ้น กว่าการตัดคำแบบอาศัยกฎ

4.2.1.3 หลักการตัดคำแบบอาศัยพจนานุกรม

หลักการตัดคำแบบอาศัยพจนานุกรม นั้น คือ การเก็บคำภาษาไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้าไปค้นหาและเทียบสายอักขระกับในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรตัดคำบริเวณใด โดยเทคนิคการตัดคำแบบอาศัยพจนานุกรมนี้สามารถตัดคำได้ถูกต้องมากกว่าการใช้กฎ ดังแสดงในงานวิจัยดังต่อไปนี้

- งานวิจัยของสมปรารถนา รัทนานนท์ (2535) ได้เสนอกลวิธีในการตัดคำภาษาไทย โดยใช้การค้นหาคำจากข้อความที่ป้อนเข้าไปเทียบกับคำในพจนานุกรม

งานวิจัยของ ดวงแก้ว สวามิภักดิ์ (2541) โดยนำพจนานุกรมมาช่วยในการจัดเก็บคำ สำหรับตรวจสอบในการตัดแยกคำ โดยผลการตัดคำในงานวิจัยนี้สามารถให้ค่าความถูกต้องในการตัดแยกคำถึง 98.11 %

- งานวิจัยของ ยืน ภู่วรรณ (2535) โดยนำพจนานุกรมมาช่วยในการจัดเก็บคำสำหรับตรวจสอบในการตัดแยกคำ และมีการถอยกลับเมื่อพบคำกำกวม โดยผลการตัดคำในงานวิจัยนี้สามารถให้ค่าความถูกต้องในการตัดแยกคำถึง 99.00 %

เทคนิคการตัดคำแบบอาศัยพจนานุกรมนี้สามารถตัดคำได้ถูกต้องมากกว่าการใช้กฎ เพราะฉะนั้นจึงเทคนิคการตัดคำที่ได้รับความนิยมสูงในปัจจุบัน

จากเทคนิคการตัดคำทั้ง 3 ที่ผ่านมา วิรัช (2538) นำเสนอถึงค่าความถูกต้องในการตัดแยกคำด้วยวิธีการต่าง ๆ โดยเป็นวิธีการตัดแยกคำที่ได้กล่าวในบทที่ 2 และเป็นวิธีการตัดคำที่อาศัยเทคนิคการตัดคำแบบพจนานุกรม และ คลังข้อมูล ช่วยในการตัดคำ จะเห็นได้ว่าการตัดคำด้วยเทคนิคพจนานุกรมและคลังข้อมูลนั้น สามารถทำให้ตัดแยกคำได้ถูกต้องและมีค่าความแม่นยำสูง ดังนี้

- Longest matching (92%)
- Maximal matching (93%)
- POS tri-gram (96%)

- Machine learning (97%)

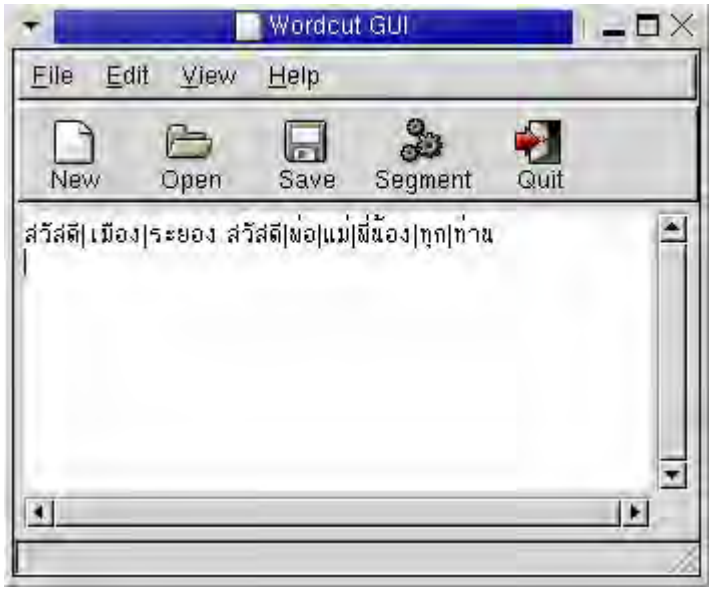
จากผลการเปรียบเทียบเห็นได้ว่าเทคนิคการตัดคำแบบอาศัยพจนานุกรมและคลังข้อมูลในการตัดแยกคำนั้น สามารถให้ค่าความถูกต้องในการตัดคำสูง ผู้วิจัยจึงได้นำมาพิจารณาเป็นเทคนิคในการตัดคำของงานวิจัย

4.2.2 โปรแกรมตัดแยกคำไทย

จากหลักการต่าง ๆ จึงได้มีการพัฒนาโปรแกรมตัดแยกคำไทย โดยโปรแกรมที่มีผู้นิยมใช้นั้น มีด้วยกันดังนี้ จึงได้มีการพัฒนาโปรแกรมตัดแยกคำที่หลากหลายรูปแบบ อาทิเช่น

4.2.2.1 โปรแกรมตัดแยกคำแบบ ThaiWordseg

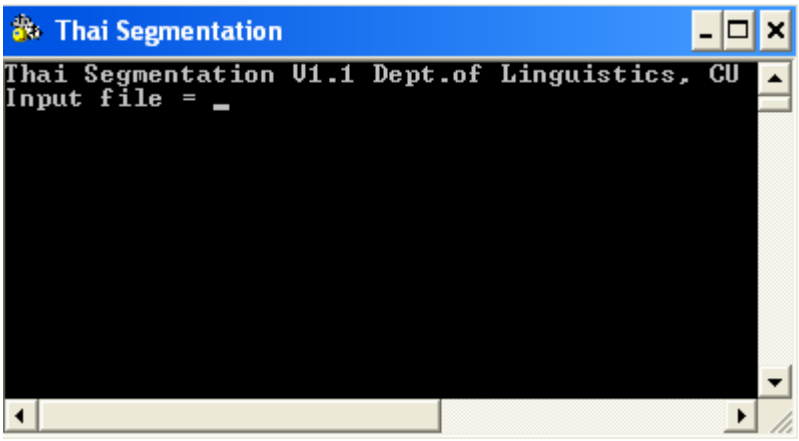
โปรแกรมการตัดคำนี้ พัฒนาโดย Vee Satayamas และ Warat Yingsalee โดยเป็นโปรแกรมตัดแยกคำที่ใช้กฎในการตัดแยกคำ โดยทำการกำหนดกฎในส่วนของโปรแกรมต้นฉบับ (source code) ของตัวโปรแกรม ด้วยการการอาศัยเทคนิคการตัดคำแบบใช้กฎนี้ อาจมีข้อจำกัดในการเพิ่มกฎในกรณีในเจอเอกสารรูปแบบที่แตกต่างไปจากเงื่อนไขที่กำหนดไว้



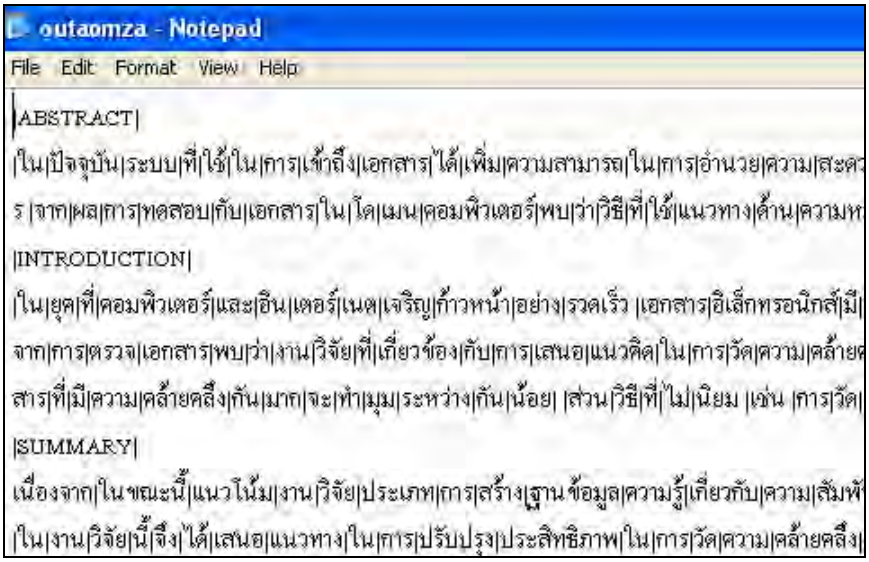
ภาพประกอบที่ 5-4 ตัวอย่างโปรแกรมตัดคำ ThaiWordSeg

4.2.2.2 โปรแกรมตัดแยกคำ Thai Word Segmentation (Version 1.1)

โปรแกรมตัดแยกคำนี้ พัฒนาโดย Wirote Aroonmanakun โดยเป็นโปรแกรมที่อาศัยพจนานุกรมข้อมูลและคลังข้อมูล รวมถึงไตรแกรมโมเดลในการตัดแยกคำ เพื่อช่วยในการแก้ไขปัญหาความกำกวมของการตัดแยกคำ แต่ข้อด้อยของโปแกรมตัดแยกคำนี้ คือ ใช้เวลาในการประมวลผลเนื่องจากพจนานุกรมมีขนาดใหญ่ถึง 630,000 syllables และสาเหตุสำคัญที่ผู้วิจัยไม่เลือกใช้โปรแกรมตัดแยกคำนี้ คือ ไม่สามารถตัดคำแบบกำกับหน้าที่คำได้ ซึ่งหน้าที่ของคำนั้นเป็นคุณสมบัติคำที่ผู้วิจัยเลือกที่จะนำมาเป็นค่าน้ำหนักของคำหรือวลีในงานวิจัยนี้



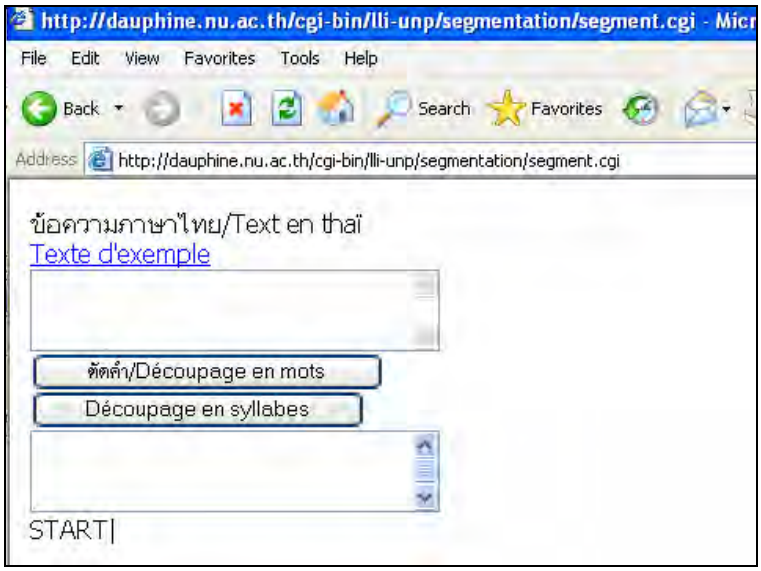
ภาพประกอบที่ 5-5 ตัวอย่างโปรแกรมตัดคำ Thai Word Segmentation (Version 1.1)



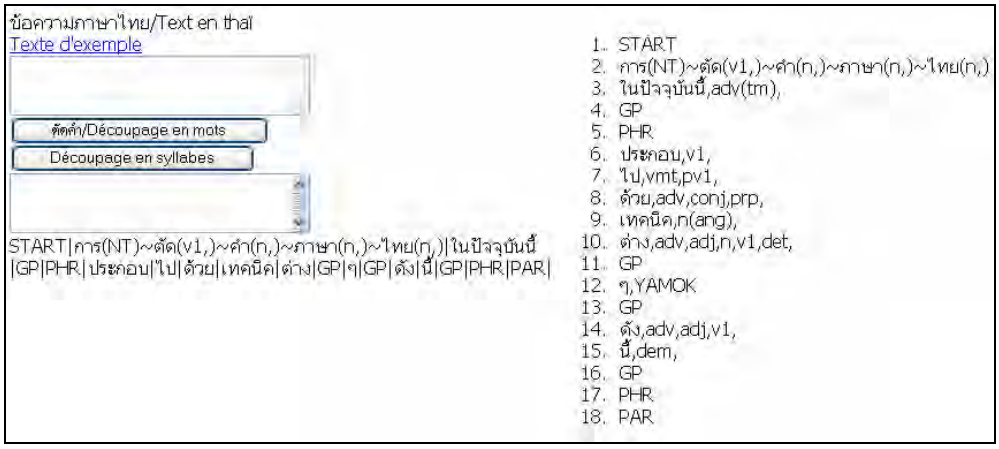
ภาพประกอบที่ 5-6 ผลการตัดแยกคำโดยโปรแกรมตัดคำ Thai Word Segmentation (Version 1.1)

4.2.2.3 โปรแกรมตัดแยกคำไทยแบบออนไลน์

โปรแกรมตัดแยกคำนี้ พัฒนาโดย โครงการ “โดฟิน” มีวัตถุประสงค์หลักเพื่อสร้างไมโครซิสเต็มด้านฐานความรู้เกี่ยวกับภาษาและวัฒนธรรม มีผู้รับผิดชอบโครงการ คือ รองศาสตราจารย์ สมบัติ เครือทอง โดยโปรแกรมตัดคำออนไลน์นี้ มีข้อดีคือเป็นการตัดคำพร้อมกำกับหน้าที่คำ แต่มีข้อด้อยคือเป็นการตัดคำแบบออนไลน์ ซึ่งอาจลำบากต่อการใช้งาน และมีจำนวนจำกัดในการตัดแยกคำต่อครั้งของการใช้งาน



ภาพประกอบที่ 5-7 ตัวอย่างโปรแกรมตัดคำแบบออนไลน์



ภาพประกอบที่ 5-8 ตัวอย่างผลของการตัดคำแบบกำกับหน้าที่ของโปรแกรมตัดคำออนไลน์

4.2.2.4 โปรแกรมตัดแยกคำ SWATh

โดยโปรแกรมตัดแยกคำ SWATh นั้น พัฒนาโดย ไพศาล เจริญพรสวัสดิ์ (2547) และเป็นโปรแกรมสำหรับการตัดแยกคำไทย โดยในงานด้านการประมวลผลภาษาธรรมชาติได้มีผู้นำโปรแกรมตัดแยกคำนี้ มาใช้ในงานวิจัยเป็นจำนวนมาก ดังนี้

- งานวิจัยของ Chuleerat Jaruskulchai , Jesada Kuntasena และ Staporn Kiewsuwansak (2544) นั้น ได้นำโปรแกรมตัดแยกคำ SWATh มาใช้ตัดแยกคำในเบื้องต้น ก่อนนำกลุ่มคำนั้นไปกำหนดค่าน้ำหนักเพื่อจัดกลุ่มเอกสาร
- งานวิจัยของ อรุณี โอฬารานนท์ (2546) โดยเป็นงานวิจัยเกี่ยวกับระบบตอบคำถามเป็นภาษาธรรมชาติ โดยได้นำโปรแกรมตัดแยกคำ SWATh มาใช้ในการตัดแยกคำออกจากประโยคที่ผู้ใช้ทำการป้อนเข้าสู่ระบบตอบคำถาม

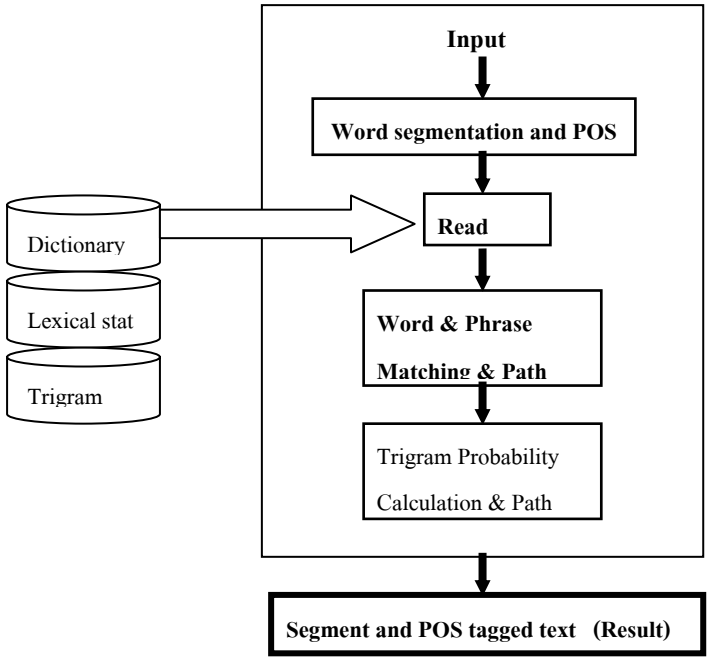
โดยโปรแกรมตัดแยกคำ SWATh นั้น ผู้วิจัยเลือกที่จะนำมาใช้ในการตัดแยกคำในเบื้องต้นของการประมวลผลในงานวิจัยนี้ โดยโปรแกรมตัดแยกคำนั้น อาศัยหลักการตัดคำแบบอาศัยพจนานุกรมและอาศัยคลังข้อมูล และภายในโปรแกรมนั้นมีเทคนิคภายในโปรแกรมให้ผู้ใช้ได้เลือกตัดแยกคำได้ถึง 3 แบบด้วยกัน คือ การตัดคำแบบยาวที่สุด การตัดคำโดยเลือกแบบเหมือนมากที่สุด และการตัดคำแบบอาศัยหน้าที่ (ไบแกรมโมเดล) โดยเป็นการอาศัยคลังข้อมูลคำศัพท์และวิธีการทางสถิติเพื่อบ่งบอกถึงความหมายของคำ

โปรแกรมตัดแยกคำ SWATh นั้น มีข้อดีสำหรับบงานวิจัย ดังนี้

1. อาศัยเทคนิคการตัดคำแบบพจนานุกรมและคลังข้อมูลช่วยในการตัดแยกคำทำให้ค่าความถูกต้องในการตัดคำนั้น มีค่าความถูกต้องสูง
2. เป็นการทำงานที่อาศัยวิทยาการศึกษาสำนึกและวิธีการทางสถิติเข้ามาช่วยแก้ปัญหาความกำกวมในการตัดคำ
3. เหมาะสำหรับการตัดแยกคำที่เป็นเอกสารภาษาไทย
4. โครงสร้างพจนานุกรมของโปรแกรมนั้น อาศัยโครงสร้างข้อมูลแบบทรัย[1] สำหรับใช้ในการจัดเก็บข้อมูลคำศัพท์ทำให้สามารถค้นหาข้อมูลได้รวดเร็วยิ่งขึ้น
5. ในการตัดคำนั้นงานด้านประมวลผลภาษาธรรมชาติต่าง ๆ นั้น นอกจากจะต้องการรู้ของเขตของคำแล้วยังจำเป็นที่จะต้องทราบ “ หน้าที่คำ ” หรือ ความหมายของคำด้วยเพื่อที่จำสามารถนำไปใช้ในบ่งบอกถึงชนิดของคำนั้น ๆ ได้ดียิ่งขึ้น

ซึ่งผู้วิจัยได้เล็งเห็นถึงความสำคัญของหน้าที่คำทางไวยากรณ์ภาษา รวมถึงโปรแกรม SWATh ที่นำมาใช้นั้น ได้อาศัยวิธีการทางสถิติ คือ การใช้ค่าสถิติที่เกิดจากลำดับหน้าที่คำหรือไวยากรณ์ทางภาษาด้วย ดังนั้นในงานวิจัยนี้ผู้วิจัยจึงเลือกที่จะนำหน้าที่ของคำมากำหนดเป็นคุณสมบัติหนึ่งของค่าน้ำหนักคำหรือวลีด้วย ผู้วิจัยจึงเลือกใช้โปรแกรมตัดแยกคำ SWATh พร้อมทั้งเลือกตัดแยกคำแบบกำกับหน้าที่ ไบแกรม ที่สอดคล้องกับไวยากรณ์ภาษา

4.2.3 กระบวนการการตัดคำและกำกับหน้าที่คำ



ภาพประกอบที่ 5-9 ขั้นตอนการตัดคำและกำกับหน้าที่คำ

ดังภาพประกอบที่ 5-9 นั้น มีรายะเอียดของแต่ละขั้นตอนดังนี้

4.2.1.1 ขั้นตอนการเลือกระบุการป้อนแอกสาร .txt หรือ .doc เข้าสู่โปรแกรมตัดแยกคำ

4.2.1.2 ขั้นตอนการตัดแยกคำพร้อมทั้งกำกับหน้าที่คำ

4.2.1.3 ขั้นตอนการอาศัยหลักสถิติเปรียบเทียบการใช้คำ กับพจนานุกรมของโปรแกรมสถิติการใช้คำร่วมกันในเอกสาร และ สถิติไตรแกรมโมเดล

4.2.1.4 ขั้นตอนการตรวจสอบคำกับพจนานุกรมคลังคำศัพท์

4.2.1.5 ขั้นตอนตอนการเปรียบค่าที่เป็นไปได้มากที่สุดของการเปรียบเทียบคำที่ตัดกับคำในพจนานุกรมคลังคำศัพท์

การตัดคำโดยใช้หน้าที่คำแบบไตรแกรมโมเดล คือ การตัดคำโดยมีการนำเอาค่าสถิติ ซึ่งพิจารณาจากความต่อเนื่องของหน้าที่คำ ส่วนวิธีการเลือกแบบตัดคำที่ดีที่สุดนั้นได้โดยหาประโยคที่มีความน่าจะเป็นมากที่สุด โดยการหาความน่าจะเป็นของแต่ละประโยค ดังที่ได้กล่าวไว้ในบทที่ 2

จากขั้นตอนที่กล่าวมานี้เป็นภาพรวมของการทำงานของโปรแกรมตัดแยกคำ โดยรายละเอียดของการเรียกใช้งานโปรแกรมและชนิดการเลือกตัดแยกคำนั้น จะกล่าวถึงในหัวข้อต่อไป

4.2.2 ขั้นตอนการตัดคำและกำกับหมวดคำในคลังข้อมูลฝึกสอน

การตัดคำและกำกับหมวดคำภาษาไทยให้กับแต่ละคำหรือวลีในเอกสารข้อมูลฝึกสอนและเอกสารข้อมูลทดสอบในงานวิจัยนั้น ผู้วิจัยเลือกใช้โปรแกรมตัดแยกคำ SWATh และเลือกที่จะทำการตัดคำแบบกำกับหน้าที่คำที่สอดคล้องกับหลักไวยาณ์ภาษา โดยข้อมูลภาษาไทยที่ทำการตัดคำและกำกับหมวดคำแล้วจะมีรูปแบบดังภาพประกอบที่ 5-10

```
ฉัน@NCMN/เดิน@NPRP/ไป@VCRJ/โรงเรียน@NCMN
```

ภาพประกอบที่ 5-10 ตัวอย่างของการกำกับหมวดคำ หลังจากผ่านการตัดคำ

4.2.2.1 คำสั่งในการตัดคำคำแบบหน้าที่คำ มีขั้นตอนดังนี้

```
swath –m bip < inputfile.txt > outputfile.txt
(ดังภาคผนวก ง.)
```

ภาพประกอบที่ 5-11 ตัวอย่างของคำสั่งในการตัดแยกคำในงานวิจัย

โดยมีรายละเอียดของคำสั่งในการตัดแยกคำ ดังนี้

swath เป็นการเรียกใช้งาน SWATh

-m เป็นการเรียกใช้งานอัลกอริทึมในการตัดคำแบบใด

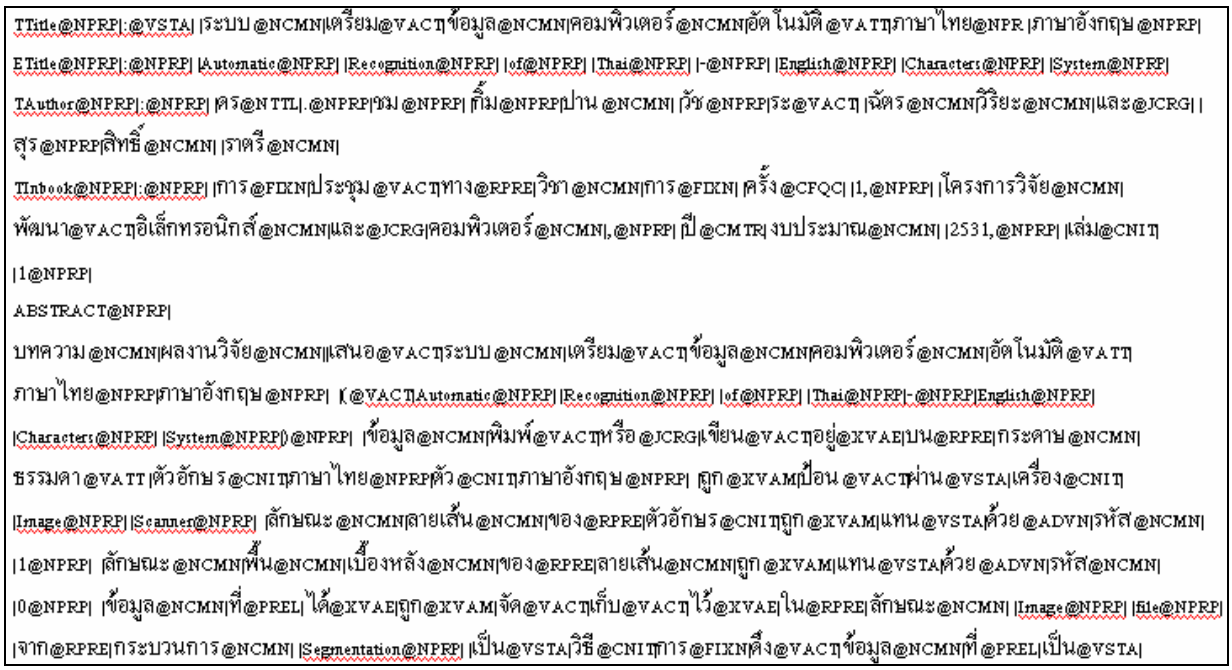
bip เป็นการเลือกใช้อัลกอริทึมในการตัดคำแบบ ไบแกรม และมีหน้าที่คำ (ไบแกรม algortihm with part-of-speech tag)

<inputfile.txt> เป็นชื่อของไฟล์เอกสารที่จะใช้ในการตัดคำ

Outputfile.txt เป็นการกำหนดชื่อไฟล์เอกสารที่ต้องการให้โปรแกรมนำผลลัพธ์จากการตัดคำที่ได้ไปแสดงผล

4.2.2.2 รูปแบบของคำหรือวลีพร้อมหน้าที่คำ

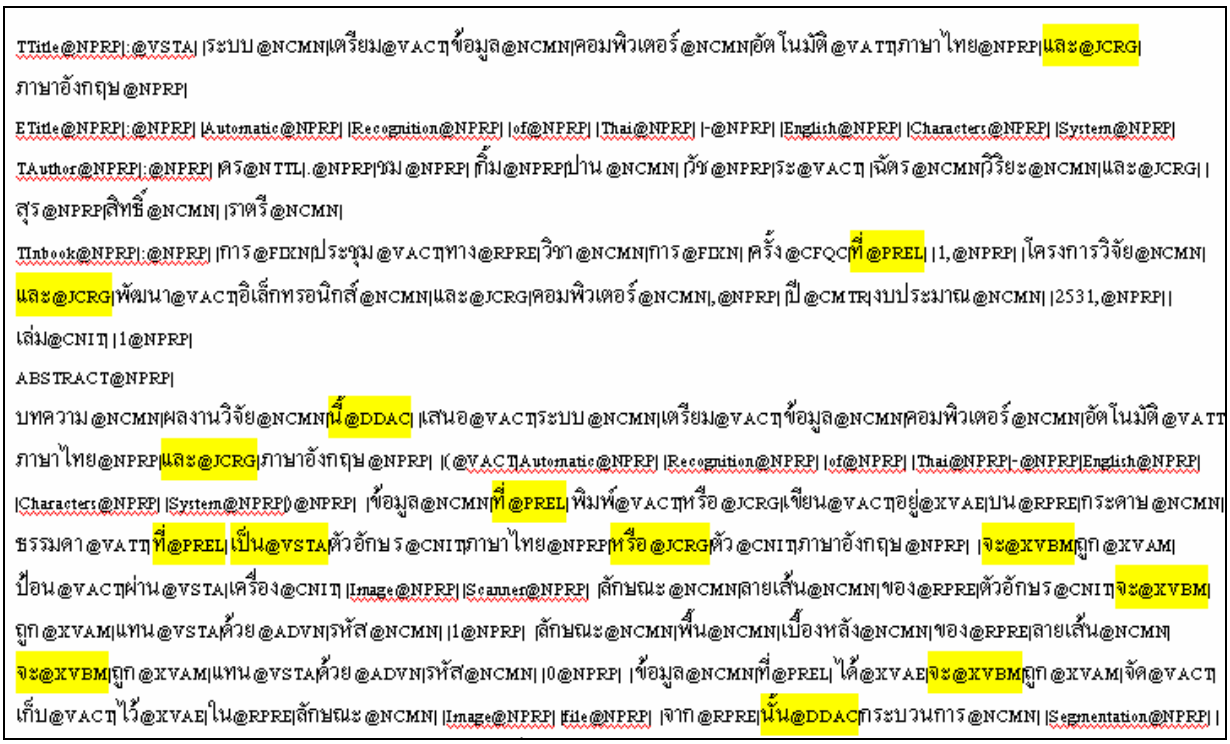
ในหัวข้อนี้จะเป็นการแสดงถึงผลลัพธ์ของการตัดแยกคำหรือวลี หลังจากผ่านการตัดแยกคำโดยโปรแกรมตัดแยกคำ โดยจะได้ไฟล์ที่มีลักษณะแยกเป็นคำคั่นด้วยเครื่องหมาย “ ไปป์ ” (|) เพื่อบ่งบอกคำแต่ละคำ และในแต่ละคำนั้นจะมีการกำกับคำด้วยหน้าที่ของคำ ดังที่ได้กล่าวมาในหัวข้อก่อนหน้านี้ โดยคำแต่ละคำในเอกสารทดสอบจะถูกตัดและแปลงให้กำกับด้วยสัญลักษณ์หมวดคำโดยใช้เครื่องหมาย “ แอท ” หรือ “ อาร์รอบ” (arrobe) (@) คั่นระหว่างรูปคำและสัญลักษณ์หมวดคำ



ภาพประกอบที่ 5-12 รูปแบบการตัดคำโดยโปรแกรม SWATh ที่ยังไม่มีการกลั่นกรองคำหยุด

4.3 การออกแบบการกลั่นกรองคำหยุด

คำหยุด นั้นเป็นคำที่เกิดขึ้นเป็นจำนวนมากในเอกสาร และไม่มีความสำคัญที่จะนำมาเป็นตัวแทนของเอกสารได้ ผู้วิจัยจึงได้เลือกที่จะพิจารณานำคำหยุดออกจากเอกสารฝึกสอนและเอกสารทดสอบก่อนการนำไปใช้ในการเรียนรู้และทดสอบของโครงข่าย โดยกลุ่มคำหยุดที่เกิดขึ้นในเอกสารนี้นั้น ดังแสดงในส่วนที่ได้มีการระบายสีไว้ ซึ่งในงานวิจัยนี้จะมีการกลั่นกรองคำหยุดก่อนในเบื้องต้น โดยจะทำการกลั่นกรองคำหยุดหลังจากผ่านโปรแกรมการกำหนดค่าน้ำหนัก ดังที่จะกล่าวในหัวข้อ 4.5 เพื่อความเหมาะสมก่อนที่จะนำไปให้โครงข่ายประสาทเทียมเรียนรู้ และทดสอบ ดังภาพประกอบ 5-13



ภาพประกอบที่ 5-13 รูปแบบการตัดคำโดยโปรแกรมที่ยังไม่มีการกลั่นกรองคำหยุด

4.4 การออกแบบรูปแบบของข้อมูลเข้าสู่โครงข่าย (ข้อมูลค่าน้ำหนักคำหรือวลี)

หัวข้อนี้จะเป็นการกล่าวถึงรูปแบบของข้อมูลสำหรับการเรียนรู้และการทดสอบของโครงข่ายประสาทเทียม โดยเป็นข้อมูลที่อยู่ในรูปแบบของเมทริกซ์ค่าน้ำหนักของคำหรือวลี เมตริกซ์ค่าน้ำหนักที่จะนำเข้าสู่ระบบเพื่อนำมาประมวลผลนั้นจะถูกแทนด้วยตัวเลข โดยที่ในแต่ละแถวของคำหรือวลีจะประกอบไปด้วยค่าน้ำหนักของคำที่บ่งบอกถึงระดับความสำคัญของคำหรือวลีนั้น ๆ ในเอกสาร โดยจะมีลักษณะในรูปแบบของเมทริกซ์ ดังนี้ คำ (W) ความน้ำหนักความถี่ของคำที่เกิดขึ้นในเอกสาร (Freq) ค่าน้ำหนักของคำตามหน้าที่ของคำ (POS) ค่าน้ำหนักคำของตำแหน่งที่เกิดขึ้นในประโยค (PS) ค่าน้ำหนักคำของตำแหน่งที่เกิดขึ้นในย่อหน้า (PP) ค่าน้ำหนักคำของตำแหน่งที่เกิดขึ้นในเอกสาร (PD) และค่าผลการประเมินที่ระบุว่าเป็นใจความคำหรือวลีสำคัญหรือไม่ เพื่อให้ระบบโครงข่ายประสาทเทียมได้เรียนรู้และทดสอบ โดยอยู่ในรูปแบบของเมทริกซ์ขนาด n x m ดังภาพประกอบ 5-14

ระบบโครงข่ายประสาทเทียมได้เรียนรู้และทดสอบ ที่อยู่ในรูปแบบของเมทริกซ์ขนาด n x m ดังภาพประกอบ 5-14

Word & Phrase =

$$\begin{pmatrix} W(1,1) & Freq_{W(1,1)} & PS_{W(1,1)} & PP_{W(1,1)} & PD_{W(1,1)} & POS_{W(1,1)} & Key_{W(1,1)} \\ W(2,1) & Freq_{W(2,1)} & PS_{W(2,1)} & PP_{W(2,1)} & PD_{W(2,1)} & POS_{W(2,1)} & Key_{W(2,1)} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ W(n,1) & Freq_{W(n,1)} & PS_{W(n,1)} & PP_{W(n,1)} & PD_{W(n,1)} & POS_{W(n,1)} & Key_{W(n,1)} \end{pmatrix}$$

ภาพประกอบที่ 5-14 แสดงรูปแบบของคำในการเรียนรู้และทดสอบของโครงข่าย

โดยสิ่งที่แสดงอยู่ในแต่ละตำแหน่งของคู่อันดับในเมทริกซ์คำ คือ ค่าน้ำหนักของคำหรือวลี โดยรายละเอียดของค่าน้ำหนัก ดังนี้

- Freq คือ ค่าน้ำหนักความถี่การเกิดขึ้นของคำหรือวลีเดิมซ้ำ ๆ ในเอกสาร
- PS คือ ค่าน้ำหนักตำแหน่งที่เกิดขึ้นในประโยคที่มีค่ามากที่สุดของคำหรือวลี
- PP คือ ค่าน้ำหนักตำแหน่งที่เกิดขึ้นในย่อหน้าที่มีค่ามากที่สุดของคำหรือวลี
- POS คือ ค่าน้ำหนักหน้าที่คำที่มีค่ามากที่สุดของคำหรือวลีนั้น
- PD คือ ค่าน้ำหนักตำแหน่งที่เกิดขึ้นในเอกสารที่มีค่ามากที่สุดของคำหรือวลี

4.4.1 รูปแบบของข้อมูลพร้อมค่าน้ำหนัก

Freq	PS	PP	POS	PD
2	0.75	1	0.9	1
11	1	1	0.9	1
7	1	1	0.9	1
19	1	1	0.9	1
2	0.75	0.75	0.9	1

ภาพประกอบที่ 5-15 ตัวอย่างชุดของข้อมูลอินพุต 5 ค่าน้ำหนักของคำหรือวลี

Key

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

ภาพประกอบที่ 5-16 ตัวอย่างข้อมูลเอาต์พุตหรือข้อมูลเป้าหมาย (Target)

ดังภาพที่ 5-15 และ 5-16 นั้นเป็นข้อมูลค่าน้ำหนักและข้อมูลเป้าหมายของแต่ละคำหรือวลี โดยจะได้ข้อมูลเมตริกซ์ค่าน้ำหนักของคำหรือวลีที่จะป้อนเข้าสู่โครงข่ายแบบ [1 x 5] [1x1] โดย

- เมตริกซ์ขนาด [m x 5] นั้น คือ ข้อมูลค่าน้ำหนักของคำหรือวลี โดย m คือ จำนวนแถวของคำหรือวลี ดังภาพประกอบที่ 5-15
- [1x1] เมทริกซ์ของคำอ้างอิงที่สำหรับใช้ในการตรวจสอบว่าเป็นคำหรือวลีสำคัญหรือไม่ ดังภาพประกอบที่ 5-16

4.5 ขั้นตอนการให้ค่าน้ำหนักคำจากอัลกอริทึมกฎการเรียนรู้ทางไวยากรณ์ภาษาศาสตร์

การกำหนดค่าน้ำหนักของคำหรือวลีนั้น เป็นการแปลงลักษณะระดับความสำคัญของคำหรือวลีให้อยู่ในรูปแบบของค่าตัวเลขที่เหมาะสมสำหรับการเรียนรู้ของโครงข่าย เพราะโดยหลักการทำงานของโครงข่ายนั้น จะเป็นการประมวลค่าข้อมูลที่เป็นตัวเลข โดยเป็นการสร้างข้อมูลคุณสมบัติของคำหรือวลีในรูปแบบของเมตริกซ์ค่าน้ำหนักของข้อมูล ดังที่ได้กล่าวในหัวข้อที่ผ่านมา สำหรับในส่วนของขั้นตอนการกำหนดค่าน้ำหนักของคำหรือวลีนั้น จะทำโดยผ่านโปรแกรมให้ค่าน้ำหนักคำหรือวลี ที่ผู้วิจัยได้พัฒนาขึ้นด้วยภาษาซี (C) ตามหลักการของอัลกอริทึมที่อ้างอิงหลักไวยากรณ์ภาษา โดยรายละเอียดการกำหนดค่าน้ำหนักนั้น ดังนี้

4.5.1 การให้ค่าน้ำหนัก (Term Weighting)

จากการตัดแบ่งคำแบบกำกับหน้าที่คำในหัวข้อที่ผ่านมานั้น เพื่อให้ได้กลุ่มคำหรือวลีที่มีการกำกับหน้าที่คำนั้น สามารถนำเข้าสู่โปรแกรมให้ค่าน้ำหนักได้ดังนี้

1. อ่านข้อมูลเข้ามาทีละอักขระ
2. ตรวจสอบว่าอักขระที่อ่านมาเป็นสัญลักษณ์ที่แสดงถึงย่อหน้าหรือไม่ ถ้าเป็น ย่อหน้าใหม่ ให้เริ่มต้นนับค่าที่บอกถึงตำแหน่งของบรรทัดในย่อหน้าใหม่
3. ตรวจสอบว่าอักขระที่อ่านมาเป็นสัญลักษณะที่แสดงถึงการสิ้นสุดบรรทัดหรือไม่ ถ้าเป็นการสิ้นสุดบรรทัด ให้เพิ่มค่าที่บอกถึงตำแหน่งของบรรทัดในย่อหน้าใหม่
4. ถ้าเจอ อักขระ “@” ต่อจากอักขระนี้จะเป็นหน้าที่คำของแต่ละคำหรือวลีนั้น ๆ ทำการอ่าน 4 อักขระของหน้าที่คำจนหมด จัดเก็บลงตารางชั่วคราวของหน้าที่คำ
5. ถ้าเจอ อักขระ “|” ถือว่าสิ้นสุดแต่ละคำหรือวลี ดังนั้นทำการเพิ่มค่าความถี่การจำนวนคำในเอกสาร
6. ตรวจสอบค่าน้ำหนักหน้าที่คำของคำหรือวลีนั้น ๆ จากตารางชั่วคราวเก็บหน้าที่คำจากนั้นทำการกำกับน้ำหนักหน้าที่คำเทียบกับตารางค่าน้ำหนักหน้าที่คำ
7. ตรวจสอบตำแหน่งการเกิดขึ้นของคำหรือวลีในประโยค และทำการกำกับค่าน้ำหนักเทียบกับตารางค่าน้ำหนักตำแหน่งในประโยค
8. ตรวจสอบตำแหน่งการเกิดขึ้นของคำหรือวลีในย่อหน้านั้นๆ ว่าเป็นคำหรือวลีที่เกิดขึ้นในบรรทัดใดของย่อหน้า ดังกฎต่าง ๆ ในการให้ค่าน้ำหนัก[1] และทำการกำกับค่าน้ำหนักเทียบกับตารางค่าน้ำหนักตำแหน่งในย่อหน้า
9. ตรวจสอบตำแหน่งการเกิดขึ้นของคำหรือวลีในเอกสาร ว่าเป็นส่วนใดของเอกสารนั้น ๆ ดังกฎต่าง ๆ ในการให้ค่าน้ำหนัก[1] และทำการกำกับค่าน้ำหนักเทียบกับตารางค่าน้ำหนักตำแหน่งในเอกสาร
10. หลังจากทำการกำกับค่าน้ำหนักให้กับแต่ละคำหรือวลีจนครบทุกค่าน้ำหนักแล้วนั้น

    10.1 ตรวจดูว่าคำนั้นเป็นคำซ้ำหรือไม่ (มีอยู่ในตารางชั่วคราวหรือไม่)

    10.2 ถ้าคำนั้นเป็นคำที่เคยอ่านเข้ามาแล้ว (เป็นคำซ้ำ)

    - ทำการเพิ่มค่าความถี่คำซ้ำของคำนั้น ๆ
    - เปรียเทียบค่าน้ำหนักในแต่ละค่าทั้ง 5 ค่าน้ำหนัก ว่าในแต่ละค่าน้ำหนักนั้น ค่าน้ำหนักปัจจุบันกับค่าน้ำหนักก่อนหน้า(ล่าสุด)ของคำนั้นที่จัดเก็บไว้ ค่าน้ำหนักใด

---

[1] ดูกฎการให้ค่าน้ำหนัก ในบทที่ 4 หน้า 78-80

มีค่ามากกว่ากัน ให้เอาค่าน้ำหนักที่มากที่สุดของแต่ละค่าน้ำหนักนั้น จัดเก็บแทนที่ในค่าน้ำหนักปัจจุบันของคำหรือวลีนั้นในตารางจัดเก็บ

10.3 ถ้าคำนั้นยังไม่เคยเข้ามา ให้เริ่มจัดเก็บเป็นคำใหม่

11. วนซ้ำอ่านมาจนหมดเอกสาร

12. ทำการตรวจสอบคำที่มีทั้งหมดในฐานข้อมูลจากไฟล์ sw.txt (stopword.txt) กับข้อมูลคำหยุดในไฟล์คำหยุด ถ้าเจอกับที่อ่านมาเป็นคำหยุดให้ทำการเอาออกจากผลการกำกับค่าน้ำหนัก เพื่อไม่เอาไปคิดน้ำหนัก

13. เขียนค่าน้ำหนักในรูปแบบของเมตริกซ์ของคำหรือวลี เพื่อส่งต่อให้กับโครงข่ายประสาทเทียมทำการทดสอบต่อไป

4.6 การออกแบบที่เกี่ยวข้องกับโครงข่ายประสาทเทียม

โดยจุดมุ่งหมายของการสอนโครงข่ายประสาทเทียมหรือการป้อนข้อมูลที่ต้องการให้กับคอมพิวเตอร์เรียนรู้ คือ การทำให้ระบบคอมพิวเตอร์นี้สามารถแสดงคำตอบในรูปแบบที่ต้องการได้โดยในงานวิจัยนี้คือ “ การสกัดคำหรือวลีสำคัญ ” ซึ่งการป้อนข้อมูลอินพุตนั้นจะเป็นการป้อนข้อมูลที่ถือว่ารู้อยู่แล้วเข้าไปให้โครงข่ายประสาทเทียมหรือข้อมูลชุดสอน พร้อมทั้งข้อมูลเป้าหมายหรือค่าเอาต์พุตที่ต้องการให้โครงข่ายประสาทเทียมนั้นแสดงออกมา โดยเป็นข้อมูลที่บอกว่าคำนั้นเป็นคำสำคัญหรือไม่ จากนั้นโครงข่ายประสาทเทียมจะทำการคำนวณและปรับค่าตัวเลขน้ำหนักเองโดยใช้กฎเกณฑ์ต่าง ๆ เข้าช่วย จนกระทั่งเอาต์พุตที่ได้ออกมานั้นถูกต้องแม่นยำอยู่ในเกณฑ์ที่น่าพอใจ ส่วนตัวเลขที่เป็นกลไกที่ทำให้โครงข่ายประสาทเทียมสามารถเรียนรู้และจดจำได้ในงายวิจัยนี้ นั้น คือ “ เมตริกซ์ของคำและค่าน้ำหนักของแต่ละคำ ”

โครงข่ายประสาทเทียมมีคุณสมบัติที่ดีและเหมาะสมต่อการใช้ในงานวิจัยนี้ ดังนี้

1. เหมาะสมสำหรับความต้องการในการดึงโครงสร้างออกจากข้อมูลที่มีอยู่ ซึ่งในงานวิจัยนี้เราต้องการดึงหรือทราบถึงโครงสร้างความสัมพันธ์ของเนื้อหาหรือข้อมูลในเอกสารในเรื่องของตำแหน่งและความถี่ของวลีสำคัญ ณ ตำแหน่งนั้น ๆ ในแต่ละประโยค และแต่ละส่วนของเอกสาร

2. สามารถสร้างแบบจำลองสำหรับงานที่ซับซ้อนมากๆได้ เช่น แบบจำลองแสดงความสัมพันธ์หลายตัวแปร การประเมินความสำคัญของวลีในงานวิจัยนี้พิจารณาจาก

ความสัมพันธ์ของตัวแปรหลายตัว คือ ค่าความสำคัญของวลีที่ชื่อเรื่อง บทคัดย่อ เนื้อความ บทสรุป และจำนวนคำในวลี

3. เหมาะกับปัญหาการแยกประเภทข้อมูล (classification problem) การพิจารณาเลือกวลีสำคัญในเอกสารจัดเป็นปัญหาการแยกประเภทข้อมูล โดยแบ่งวลีทั้งหมดในเอกสารเป็น 2 กลุ่ม คือวลีสำคัญ และวลีทั่วไป

4. สามารถทำงานกับข้อมูลที่มีความผิดพลาด หรือข้อมูลที่ไม่มีโครงสร้างได้ ซึ่งเป็นลักษณะของข้อมูลประเภทเอกสาร

5. สามารถตอบคำถามที่ไม่เคยถูกถามมาก่อนได้ ทำให้แบบจำลองโครงข่ายประสาทเทียมสำหรับเลือกวลีสำคัญที่สร้างโดยการเรียนรู้จากเอกสารตัวอย่าง สามารถนำมาใช้เลือกวลีสำคัญจากเอกสารทดสอบได้

4.6.1 การเลือกใช้โครงข่ายประสาทเทียม

โดยในงานวิจัยนี้ เลือกใช้โครงข่ายประสาทเทียมประสาทเทียมแบบมีผู้สอน โดยใช้ข้อมูลแบ่งออกเป็นข้อมูลอินพุตและข้อมูลเป้าหมายในการฝึกสอน และหลังจากที่โครงข่ายได้รับการฝึกสอนแล้วนั้น ก็จะเป็นเข้าสู่ขั้นตอนของการทดสอบพื่อคำนวณหาผลว่าคำหรือวลีใดที่มีความสำคัญของเอกสาร โดยรายละเอียดของโครงข่ายประสาทเทียมที่เลือกใช้มีด้วยกันนี้

4.6.1.1 การเลือกใช้ชนิดของโครงข่าย

โดยในงานวิจัยนี้เลือกใช้โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น หรือโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ซึ่งเป็นโครงข่ายประสาทเทียมแบบต้องมีผู้สอน และเป็นโครงข่ายประเภทการคาดเดา ที่มักใช้ทำงานกับงานคาดเดาสามารถช่วยกำหนดลำดับความสำคัญได้ โดยในงานวิจัยนี้นำมาใช้สำหรับการคาดเดาและทำนายว่าคำหรือวลีนั้นเป็นคำหรือวลีสำคัญหรือไม่

4.6.1.2 การเลือกใช้วิธีการเรียนรู้

โดยขั้นตอนการเรียนรู้ลักษณะของใจความสำคัญของคำหรือวลี จะใช้แบ็กพรอพาเกชั่นโครงข่ายประสาทเทียม ซึ่งเป็นวิธีการเรียนรู้ที่เหมาะสมกับการนำไปใช้ในการแก้ปัญหาที่เกี่ยวข้องกับการค้นหาสิ่งที่ซ่อนอยู่ซึ่งเป็นคุณสมบัติของใจความสำคัญของคำหรือวลีสำคัญในการเรียนรู้คุณสมบัติของคำหรือวลีสำคัญในเอกสารชุดสอน และสถาปัตยกรรม

แบ็คพรอพาเกชั่น นี้เป็นที่นิยมสูงสุด เนื่องจากมีประสิทธิภาพและง่ายที่จะเป็นต้นแบบสำหรับโครงข่ายประสาทเทียมที่มีความซับซ้อน

4.6.1.3 การเลือกใช้ฟังก์ชันการสอน

โดยในงานวิจัยนี้ได้เลือกใช้ฟังก์ชั่นการสอน แบบ Trainlm (Levenberg-Marquardt backpropagation) เนื่องจากมีความเร็วในการประมวลผลสูง

4.6.1.14 การเลือกใช้ฟังก์ชันปรับการเรียนรู้

โดยในงานวิจัยนี้ได้เลือกใช้ฟังก์ชันการปรับการเรียน (adaptation learning function) แบบ Learngdm ในการปรับค่าการเรียนรู้ของโครงข่าย

4.6.1.5 การเลือกใช้ฟังก์ชันทรานเฟอร์

โดยในงานวิจัยนี้ได้เลือกใช้ทรานส์เฟอร์ฟังก์ชันที่ใช้ในการกระตุ้นและปรับและถ่ายโอนแบบ Logsig นี้ด้วย โดยเพราะผลลัพธ์ของงานวิจัยมีผลลัพธ์หรือเอาต์พุต 2 ค่าที่เป็นไปได้ คือ 0 และ 1 โดย 0 คือ ไม่ได้เป็นคำหรือวลีสำคัญ และ 1 คือ เป็นคำหรือวลีสำคัญ

$$logsig(n) = 1/(1+exp(-n)) \quad ...(5\text{-}1)$$

4.6.1.6 การเลือกใช้ เทคนิคการคำนวณค่าความผิดพลาด

โดยในงานวิจัยนี้เลือกใช้เทคนิคการคำนวณค่าความผิดพลาดเฉลี่ย MSE ในการกำหนดความคลาดเคลื่อนในการทำนาย ในกรณีที่ค่าความผิดพลาดเฉลี่ยมีค่าน้อยกว่าที่ยอมรับได้ ให้จบการเรียนรู้

4.6.1.8 การเลือกใช้เครื่องมือโครงข่ายประสาทเทียม

โดยในงานวิจัยนี้เลือกใช้ Neural Network Toolbox ของโปรแกรม Matlab เพราะง่ายต่อการใช้งานโดยมีฟังก์ชันที่สามารถเรียกใช้งานโดยผ่านการเขียนคำสั่ง Command Line เพิ่มเติมได้ ดังภาคผนวก จ. และเป็นภาษาโปรแกรมชั้นสูงที่ใช้ควบคุมลำดับการทำงาน ลักษณะการเขียนโปรแกรมเป็นแบบออบเจ็กต์ (Object Oriented Programming) ทำให้การเขียนโปรแกรมไม่ยุ่งยากเมื่อเทียบกับการเขียนโปรแกรมด้วยภาษาอื่น ๆ

ข้อดีของ Neural Network Toolbox ใน Matlab

1. ง่ายต่อการใช้งาน

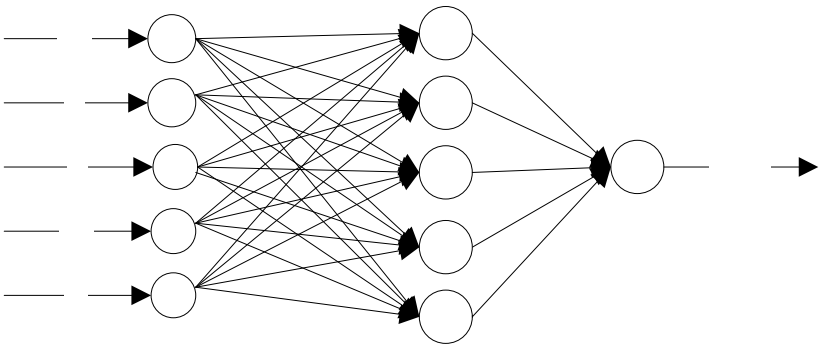
2. กล่องเครื่องมือใน Matlab นั้น ได้มีการจัดเตรียมกลุ่มของฟังก์ชันสำหรับใช้ในการจัดการโครงข่ายประสาทเทียมไว้ให้กับผู้ใช้แล้ว โดยที่ผู้ใช้ไม่จำเป็นต้องเขียนโปรแกรมต้นฉบับใหม่ รวมถึงยังมีการจัดเตรียมในส่วนของแอคติเวชั่นฟังก์ชั่น อัลกอริทึมฝึกสอนโครงข่าย สำหรับใช้งานได้อย่างสะดวกต่อผู้ใช้

3. กล่องเครื่องมือใน matlab นั้น อยู่บนพื้นฐานของ network object โดย object ประกอบด้วยข้อมูล (สารสนเทศ) เกี่ยวกับทุกสิ่งที่เกี่ยวข้องกับโครงข่ายประสาทเทียม เช่น จำนวนและโครงสร้างของแต่ละชั้น การเชื่อมต่อระหว่างชั้น

4. Matlab จะจัดเตรียม ฟังก์ชันสำหรับสร้างเครือข่ายระดับสูง เช่น newlin คือ ฟังก์ชันในการสร้าง linear layer newp คือ การสร้าง perceptron newff คือ ฟังก์ชันในการสร้าง feed-forward back-propagation ที่ง่ายต่อการเรียกใช้ และเป็นฟังก์ชันที่ผู้วิจัยเลือกใช้ในงานวิจัยฉบับนี้

4.6.2 การออกแบบโครงข่ายประสาทเทียมเพอร์เซฟตรอนแบบหลายชั้น

โดยในการออกแบบโครงสร้างของโครงข่ายประสาทเทียมนั้นแสดงดังรูปที่ 5-17



ภาพประกอบที่ 5 - 17 สถาปัตยกรรมของโครงข่ายประสาทเทียมเพอร์เซฟตรอนแบบหลายชั้น

ตัวอย่างโครงสร้างของโครงข่ายประสาทเทียมเพอร์เซฟตรอนแบบหลายชั้น ที่ใช้ในการเรียนรู้และทดสอบการสกัดคำหรือวลีสำคัญในเอกสารภาษาไทย ซึ่งมีอินพุตเป็น ค่าน้ำหนักความถี่การเกิดขึ้น (Freq) ค่าน้ำหนักตำแหน่งในประโยค (PS) ค่าน้ำหนักตำแหน่งในย่อหน้า (PP)

ค่าน้ำหนักหน้าที่คำ (POS) และค่าน้ำหนักตำแหน่งในเอกสาร (PD) โดยมีเอาต์พุตเป็น ผลการทำนายว่าคำหรือวลีนั้น ๆ มีค่าความสำคัญหรือไม่ โดยถ้าเป็นคำสำคัญจะมีค่าเอาต์พุตเป็น 1 ถ้าไม่สำคัญจะมีค่าเป็น 0

จากรูปที่ 5 -17 โครงสร้างของเน็ตเวริก์ประกอบด้วย 3 ชั้น ดังนี้

- ชั้นอินพุต

ชั้นอินพุต ประกอบด้วยจำนวนนิวรอลเท่ากับ จำนวนค่าน้ำหนักของคำหรือวลีนั้น ๆ คูณด้วยจำนวนคำหรือวลีที่มีในเอกสารนั้น (กำหนดให้เป็นค่า m) ดังนั้น กรณีคำไทยข้อมูลเข้าจะมีขนาด m * 5 นิวรอน โดยค่า m ซึ่งคือจำนวนตัวอักษรทั้งหมดที่ใช้พิจารณา และ 5 จำนวนค่าน้ำหนักของแต่ละคำหรือวลีนั้น ๆ

- ชั้นซ่อน

ชั้นซ่อน มี 1 ชั้น ได้ทำการทดลองเพื่อหาจำนวนนิวรอนที่เหมาะสม (โดยได้จากค่าที่ใช้ในการฝึกแล้วให้ผลการเรียนรู้ดีที่สุด) โดยมีทั้งสิ้น 5 นิวรอน และจำนวนชั้นซ่อน 1 ชั้นซ่อนนั้นเหมาะสมต่อการเรียนรู้แล้วสำหรับงานวิจัยนี้ เพราะข้อมูลที่ใช้ทดสอบนั้น ไม่ได้เป็นข้อมูลที่มีค่าความซับซ้อนมาก โดยหน้าที่ของโหนดในชั้นซ่อนในด้าน function approximation คือ ทำการคำนวณหาค่าประมาณที่ใกล้เคียงซึ่งจะคำนวณจากแอ็คติเวชั่นที่กำหนดและมีความเหมาะสมก่อนที่จะส่งค่าประมาณไปยังโหนดแสดงผลลัพธ์ ซึ่งจะหาค่าประมาณที่ใกล้เคียงมากที่สุด

- ชั้นเอาต์พุต

ชั้นเอาต์พุต จะมีจำนวนนิวรอนเท่ากับผลการทำนายคำหรือวลีสำคัญ คือ 1 นิวรอน โดยหน้าที่ของโหนดแสดงผลลัพธ์ในด้านฟังก์ชันที่เหมาะสม (function approximation) คือ ทำการเอาค่าประมาณที่ได้จากโหนดในชั้นซ่อนมารวมกันแล้วได้ค่าประมาณที่ใกล้เคียงกับค่าที่เหมาะสม (approximation) จริงซึ่งก็จะขึ้นกับแอ็คติเวชั่นฟังก์ชันของโหนดแสดงผลลัพธ์

4.6.3 การทำงานโครงข่ายประสาทเทียม

โดยการทำงานในส่วนของโครงข่ายประสาทเทียมเพื่อการเรียนรู้และสกัดคำหรือวลีสำคัญของโครงข่ายประสาทเทียม ในส่วนของการรับข้อมูลมีขั้นตอนทั้งหมดดังนี้

4.6.3.1 การรับข้อมูลเข้าสู่โครงข่าย

การรับข้อมูลเข้าสู่โครงข่ายนั้น จะเป็นการรับข้อมูลในรูปแบบของเมตริกซ์ตค่าน้ำหนัก โดยในเบื้องต้นนั้นข้อมูลค่าน้ำหนักที่ผ่านการตัดคำนั้น จะอยู่ในรูปแบบของ

[m x 5] แต่การรับข้อมูลเข้าสู่โครงข่ายประสาทเทียมใน Matlab นั้นจะรับข้อมูลในรูปแบบ [5 x m] โดย 5 คือ จำนวนคอลัมภ์ และ m คือ จำนวนแถวข้อมูล ดังนั้นในการนำข้อมูลเพื่อเข้าสู่การทำงานของโครงข่ายนั้น จะต้องทำการทรานสโพสก่อนโดยการเปลี่ยนข้อมูลคอลัมภ์เป็นแถวและแถวเป็นคอลัมภ์

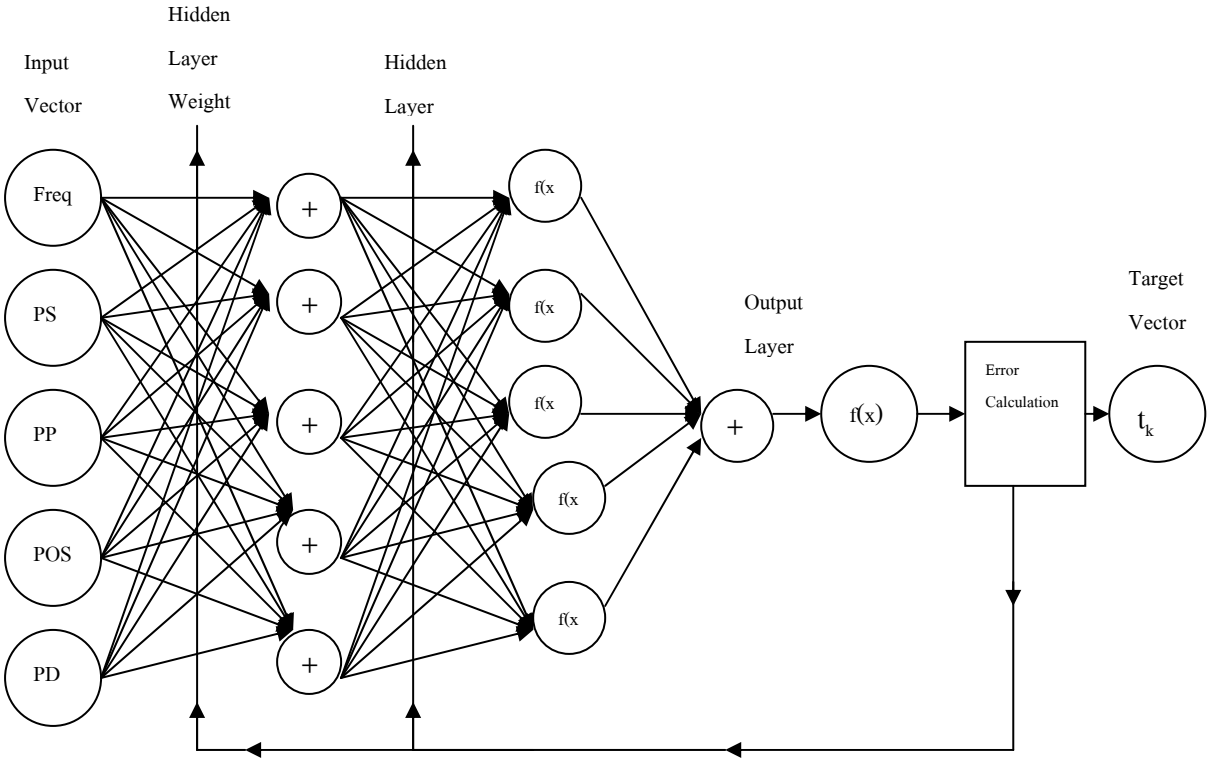
4.6.3.2 การสอนนิวรอลเน็ตเวิร์ก

ในการสอนโครงข่ายประสาทเทียมนั้นให้เรียนรู้การสกัดคำหรือวลีสำคัญนั้น ข้อมูลที่ใช้สอนได้มาจากเอกสารชุดสอนซึ่งได้มีการแปลงข้อมูลให้อยู่ในรูปแบบของค่าน้ำหนักที่เป็นตัวเลขในรูปแบบเมตริกซ์ของคำ โดยเป็นค่าน้ำหนักที่ให้ผ่านอัลกอริมกำหนดค่าน้ำหนักที่อิงกับหลักไวยากรณ์ภาษาศาตร์ของภาษาไทยดังที่กล่าวในบทที่ผ่านมา ได้แก่ หลักความถี่ หลักในการค้นหาใจความสำคัญในประโยค หลักในการค้นหาใจความสำคัญในย่อหน้า หลักในการค้นหาใจความสำคัญในเอกสาร หลักในการค้นหาใจความสำคัญของหน้าที่ของคำที่ใช้ในการเขียน หลักในการเว้นวรรคในเอกสาร หลักในการหาย่อหน้า หลักในการขึ้นต้นบรรทัด มาสร้างเป็นกฎต่าง ๆ สำหรับการกำหนดน้ำหนักของโปรแกรมกำหนดค่าน้ำหนัก

4.6.3.3 โครงข่ายประสาทเทียมในส่วนของการส่งค่าไปข้างหน้า

ในส่วนนี้จะเป็นส่วนของการทำงานในโครงข่ายประสาทเทียมแบบแพร่กลับ มีหน้าที่ในการคำนวณด้วยการคูณ การบวกและทำการส่งค่าที่ได้ผ่านฟังก์ชั่นโอนย้ายทำเช่นนี้ไปจนกระทั่งส่งค่าถึงขั้นสุดเอาต์พุตก็เป็นอันสิ้นสุดขั้นตอนการทำงานของการส่งค่าไปข้างหน้า โดยอาศัยโครงข่ายประสาทเทียมที่มีจำนวนของชั้นซ่อน 1 ชั้น และในชั้นซ่อนจะทำการทำหนดจำนวนนิวรอน 5 นิวรอนหรือ 5 โหนด ดังภาพประกอบ 5-18 โดยที่มีการรับค่าองค์ประกอบในการทำงานของนิวรอลเน็ตในช่วงของการเรียนรู้ข้อมูล คือ ค่าความผิดพลาด จำนวนชั้นซ่อน จำนวนเซลล์ในแต่ละชั้นซ่อน

ผลลัพธ์การระบุคำสำคัญหรือเอาต์พุตที่ออกจากโครงข่ายประสาทเทียมจะมีขนาด 1 เซลล์เป็นตัวเลขจำนวนจริงที่มีค่า 0 หรือ 1 ที่แสดงถึงระดับค่าความสำคัญของคำหรือวลี ว่าเป็นคำหรือวลีสำคัญหรือไม่ โดยที่ 1 แสดงถึงคำหรือวลีนั้นเป็นคำหรือวลีสำคัญในเอกสารนั้นเหมือนข้อมูลต้นแบบที่ได้เรียนรู้ ส่วนค่าเอาต์พุตที่เป็น 0 แสดงถึงว่าคำหรือวลีนั้นไม่ได้เป็นคำหรือวลีสำคัญหรือมีความสำคัญน้อยมากในเอกสาร และแสดงถึงคำหรือวลีสำคัญที่ต่างจากข้อมูลชุดสอนที่ได้เรียนรู้ไว้



ภาพประกอบที่ 5-18 แสดงรูปแบบการส่งค่าไปข้างหน้าของโครงข่ายประสาทเทียม

4.6.4 รูปแบบคำสั่ง ในการทำงานของโครงข่าย

```
net = newff(PR,[s1,s2],{TF1,TF2},BTF) ;
net = newff(minmax(Input_nn),[5,5],{'logsig','logsig'},'trainlm');
```

Command for creating the network **1**

```
PR = minmax(Input_nn);   → Range of inputs
S1  = 1-5                → No. of nodes of Layer 1
S2 = 1-5                 → No. of nodes of Layer 2
```

2

```
TF1 = 'logsig' ;         → Activation Functions of Layer 1
TF1 = 'logsig' ;         → Activation Functions of Layer 2
BTF = 'trainlm';         → Training function
```

3

ภาพประกอบที่ 5-19 แสดงคำสั่งในการออกแบบการทำงานของโครงข่าย

4.6.4.1 คำอธิบายคำสั่ง

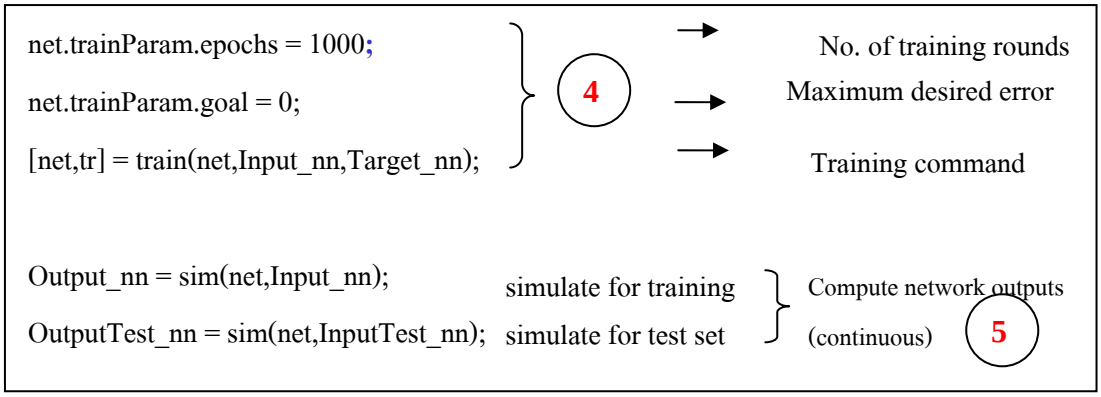
จากภาพ ภาพประกอบที่ 6-4 นั้นเป็นการอธิบายถึงรายละเอียดต่าง ๆ ของโครงข่ายที่ใช้ในการเรียนรู้และระบุวลีสำคัญ

- **1** คือ คำสั่งที่ใช้สำหรับสร้างโครงข่าย โดย net คือโครงข่ายและสถาปัตยกรรมของโครงข่าย

- **2** คือ การกำหนดในชั้นอินพุตเข้าสู่โครงข่ายและโหนดในชั้นซ่อน โดยมีจำนวนตั้งแต่ 1 ถึง 5 โหนด

- **3** คือ การกำหนดฟังก์ชันที่ใช้การเรียนรู้และปรับค่าน้ำหนักของโครงข่าย

4.6.5 รูปแบบคำสั่ง ในการฝึกสอนโครงข่าย



ภาพประกอบที่ 5-20 แสดงคำสั่งในการฝึกสอนโครงข่าย

4.6.5.1 คำอธิบายคำสั่งในการฝึกสอนโครงข่าย

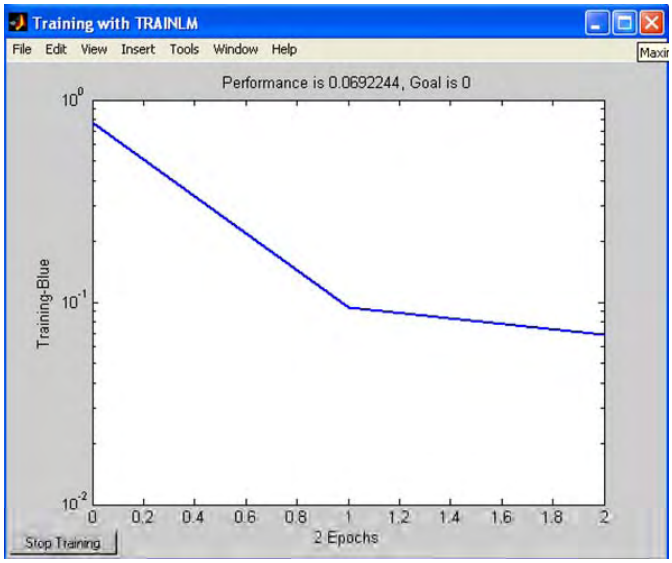
จากภาพ ภาพประกอบที่ 6-3 นั้นเป็นการอธิบายถึงรายละเอียดต่าง ๆ ของโครงข่ายที่ใช้ในการเรียนรู้และระบุวลีสำคัญ

- **4** คือ ค่าที่กำหนดจำนวนรอบในการฝึกสอนในแต่ละครั้ง คือ 1000 รอบ ค่าความผิดพลาดที่ยอมรับได้ และคำสั่งในการฝึกสอน

- **5** คือ การคำนวณผลลัพธ์ในการจำแนกของโครงข่าย

4.6.5.2 กราฟแสดงผลการเรียนรู้

ผลการเรียนจากค่าความผิดพลาดของโครงข่ายนั้น ดังภาพประกอบที่



ภาพประกอบที่ 5-21 ภาพแสดงแนวโน้มการลู่ค่าการเรียนรู้ของโครงข่าย

4.6.7 การทำนายผลและหาค่าความถูกต้องในการสกัด

เมื่อโครงข่ายได้ทำการเรียนรู้จากข้อมูลชุดฝึกสอนจนเหมาะสมแล้วนั้น ก็จะเป็นการจำแนกข้อมูลจากข้อมูลชุดทดสอบ โดยจากการทดสอบนั้น ผู้วิจัยได้กำหนดการจำแนกคำหรือวลีสำคัญจากผลการทำนายของโครงข่าย คือ ถ้าผลลัพธ์การทำนายโดยโครงข่ายมีค่า มากกว่า 0.5 นั้น ผู้วิจัยถือว่า คำหรือวลีนั้นมีความสำคัญ ดังภาพประกอบ 5-22

```
% ทำนาย class จากค่า output
for i = 1:row                    % วนเป็นจำนวนรอบเท่ากับจำนวนแถวของข้อมูลเป้าหมาย
   for j = 1:column              % ตรวจจากแถวคอลัมภ์ของค่า output target
      if  Output(i,j) >= 0.5     } % ถ้าค่าผลลัพธ์จากการเรียนรู้และทำนายของโครงข่ายได้ผลลัพธ์ที่มีค่าตั้งแต่ 0.5
         Predict(i,j) = 1;       } ขึ้นไป ถือว่าคำหรือวลีนั้นมีความสำคัญให้มีค่าความสำคัญ คือ 1
      else
         Predict(i,j) = 0;       } % แต่ถ้าค่าผลลัพธ์จากการเรียนรู้และทำนายของโครงข่ายได้ผลลัพธ์ที่มีค่าต่ำ
                                 } กว่า 0.5 ขึ้นไป ถือว่าคำหรือวลีนั้นไม่มีความสำคัญ จึงมีค่าความสำคัญ คือ 0
```

ภาพประกอบที่ 5-22 แสดงคำสั่งในการจำแนกผลลัพธ์

4.7 การออกแบบการทดสอบที่ 1

การทดสอบที่ 1 เป็นการทดสอบเพื่อประเมินว่า เมื่อจำนวนข้อมูลในอินพุต (จำนวนพารามิเตอร์ของค่าน้ำหนักคำหรือวลี) เพิ่มขึ้น เมื่อจำนวนโหนดในชั้นซ่อนที่เพิ่มขึ้น มีผลต่อค่าความถูกต้องหรือไม่ โดยการสกัดคำหรือวลีสำคัญจะเริ่มจากการฝึกสอนโครงข่ายให้สามารถเรียนรู้ถึงระดับความสำคัญของคำหรือวลี ดังที่ได้กล่าวมา และเมื่อโครงข่ายได้เรียนรู้จนเหมาะสมแล้วนั้น จึงจะเข้าสู่ขั้นตอนการทดสอบโดยชุดข้อมูลทดสอบ โดยนำผลลัพธ์ที่ได้มาเปรียบเทียบกับค่าจริงที่เกิดขึ้น โดยมีขั้นตอนการทดสอบ ดังนี้

4.7.1 การรับข้อมูล

การรับข้อมูลเมตริกซ์ที่ผ่านการกำกับค่าน้ำหนัก โดยโปรแกรมให้ค่าน้ำหนัก wordsept.exe เข้าสู่โครงข่ายโดยทำการอ่านข้อมูล ดังนี้

ข้อมูลอินพุตสอน อ่านจากไฟล์ Input_t.mat

ข้อมูลเป้าหมายสอน อ่านจากไฟล์ Target_t.mat

ข้อมูลอินพุตทดสอบ อ่านจากไฟล์ InputTest_t.mat

ข้อมูลเป้าหมายทดสอบ อ่านจากไฟล์ TargetTest_t.mat

4.7.2 การเรียกใช้งาน Neural Network Toolbox ใน Matlab 6.5

```
[net,weigth_hidden,weigth_output,acc_predict_train,acc_predict_test,bias_hidden,bias_output,OutputTest]
= train_nn_hidden_layer(Input_t,InputTest_t,Target_t,TargetTest_t,3,'logsig','logsig');
```

ภาพประกอบที่ 5-23 แสดงคำสั่งเรียกใช้งานฟังกันทำงานของโครงข่าย

คำสั่งดังภาพ 5-23 นั้นเป็นการเรียกใช้ฟังก์ชันที่ผู้วิจัยได้พัฒนาขึ้น เพื่อเรียกโครงข่ายประสาทเทียมสำหรับการฝึกสอนด้วยข้อมูลชุดสอน (Input_t.mat) และทำการทดสอบด้วยข้อมูลชุดทดสอบ (InputTest.mat) และกำหนดรายละเอียดในการทำงาน คือ และกำหนดให้โครงข่ายเรียนรู้จำนวน 1000 รอบ โดยในแต่ละเงื่อนไขการเลือกใช้พารามิเตอร์นั้น ผู้วิจัยทำการทดสอบโดยการใช้จำนวนโหนดในชั้นซ่อน ตั้งแต่ 1 ถึง 5 โหนด และในการทดสอบการ

เลือกใช้พารามิเตอร์ ต่อการเลือกใช้จำนวนโหนดในชั้นซ่อนทั้ง 5 โหนดนั้น ผู้วิจัยจะทำการทดสอบทั้งสิ้น 3 ครั้ง

4.8 การออกแบบการทดสอบที่ 2 (เทคนิคการหาค่าน้ำหนักเริ่มตอนต้นที่เหมาะสม)

ในการหาค่าน้ำหนักที่เหมาะสมเป็นค่าน้ำหนักเริ่มต้นของโครงข่าย เพื่อให้โครงข่ายมีการเรียนรู้ได้ดีที่สุดนั้น เป็นการอาศัยเทคนิคการสุ่มค่าสำหรับคูณกับข้อมูลอินพุตค่าน้ำหนักจริงของคำหรือวลี

เทคนิคการหาค่าน้ำหนักเริ่มต้นที่เหมาะสมในการป้อนเข้าสู่โครงข่ายนั้น คือ การหาค่าเริ่มต้นที่เหมาะสมกับเรียนรู้ของโครงข่าย และสามาถทำให้โครงข่ายเรียนรู้ได้ดีและสามารถจำแนกข้อมูลได้ดียิ่งขึ้น โดยโครงข่ายเพอร์เซฟตรอนแบบหลายชั้นหรือโครงข่ายแบบป้อนไปข้างหน้าและการเรียนรู้แบบแพร่ย้อนกลับนั้น คือ การเรียนรู้เพื่อปรับค่าน้ำหนักจากค่าความผิดพลาดที่เกิดขึ้น ด้วยเหตุนี้ การที่ได้ค่าที่เหมาะสมในเริ่มต้นนั้น ก็จะทำให้ค่าความผิดพลาดในการเรียนรู้นั้น สามารถที่จะลู่เข้าสู่ค่าเป้าหมายหรือในระดับกำหนดไว้ได้ดียิ่งขึ้น โดยจากงานวิจัยของ (Mercedes และ Carlos, 200) มีการกล่าวถึงเทคนิคการกำหนดค่าน้ำหนักเริ่มต้นด้วยกันถึง 7 เทคนิค ต่าง ๆ โดยมีการกล่าวถึงเทคนิคในการสุ่มค่าน้ำหนัก สำหรับสร้างค่าน้ำหนักใหม่ก่อนที่จะป้อนเข้าสู่โครงข่าย ดังงานวิจัย (Drago, G.P., Ridella, S, 1992) ได้มีการสุ่มค่าตัวเลขที่อยู่ในช่วง 0-1 โดยทำการแบ่งค่าช่วงข้อมูลที่สุ่มนั้นออกเป็น 2 ช่วง และนำไปคูณกับข้อมูลอินพุตเข้าสู่โครงข่าย

4.8.1 การสุ่มค่าน้ำหนักเพื่อสร้างค่าน้ำหนักใหม่

ในการทดสอบในการทดสอบประเภทที่ 1 นั้น ข้อมูลค่าน้ำหนักอินพุตที่ป้อนเข้าสู่โครงข่ายนั้น จะถูกคูณด้วยค่า 1 เสมอ โดยจากเทคนิคการสร้างน้ำหนักในได้กล่าวมาในตอนต้นนั้น การทดสอบประเภทที่ 2 นี้จึงเป็นการทดสอบโดยอาศัยเทคนิคการสุ่มค่าน้ำหนัก เพื่อหาค่าน้ำหนักที่เหมาะสมในแต่ละพารามิเตอร์อินพุต ที่จะทำให้ค่าความถูกต้องมีค่าสูง (ที่สุด) โดยจะเป็นวิเคราะห์ถึงความสัมพันธ์ในการให้น้ำหนักแต่ละพารามิเตอร์ ว่าควรจะกำหนดระดับความสำคัญในแต่ละคุณสมบัติมากน้อยอย่างไร สำหรับในแต่ละรูปแบบของเอกสารที่ใช้ในการสอนและทดสอบที่จะมีผลต่อค่าความถูกต้อง ดังนี้

4.8.1.1 ขั้นตอนการรับข้อมูลอินพุต

การรับข้อมูลอินพุตในการทดสอบประเภทที่ 2 นั้น จะเป็นการรับข้อมูลอินพุตที่ผ่านการตัดคำและให้ค่าน้ำหนักดังที่กล่าวในหัวข้อ 3 โดยในการทดสอบนี้จะเลือกใช้จำนวนพารามิเตอร์ 5 พารามิเตอร์ในการทดสอบ เพราะในการทดสอบนี้จะเป็นการหาค่าน้ำหนักที่เหมาะสมของแต่ละพารามิเตอร์ ในแต่ละรูปแบบของความสัมพันธ์ระหว่างเอกสารฝึกสอนและเอกสารทดสอบ ที่จะทำให้ได้ค่าความถูกต้องที่ดีที่สุด

4.8.1.2 ขั้นตอนการสร้างค่าน้ำหนักใหม่

```
% MulRandomWeight(Input_t);
SizeMatrix = size(Input_t);
RowMatrix = SizeMatrix(1,1);
ColumnMatrix = SizeMatrix(1,2);     % จำนวน column

% Multiply weight with random
   RanWeight = rand(1,5);  % random weight

   for iRow = 1:RowMatrix    % check num of read row
     %cal Column 1 with random weight 1
      RanCol1 =  RanWeight(1,1);
      InputNN(:,1) =  Input_t(:,1) * RanCol1 ;

     %cal Column 2 with random weight 2
      RanCol2 =  RanWeight(1,2);
      InputNN(:,2) =  Input_t(:,2) * RanCol2 ;

     %cal Column 3 with random weight 3
      RanCol3 =  RanWeight(1,3);
      InputNN(:,3) =  Input_t(:,3) * RanCol3 ;

     %cal Column 4 with random weight 4
      RanCol4 =  RanWeight(1,4);
      InputNN(:,4) =  Input_t(:,4) * RanCol4 ;

     %cal Column 5 with random weight 5
      RanCol5 =  RanWeight(1,5);
      InputNN(:,5) =  Input_t(:,5) * RanCol5 ;
```

การหาขนาดของจำนวนคอลัมถ์ของข้อมูลชุดฝึกสอน สำหรับใช้ในการกำหนดขนาดจำนวนคอลัมภ์ข้อมูลสุ่ม ที่จะต้องสัมพันธ์กันในการคูณค่าน้ำหนัก

การสุ่มค่าน้ำหนักตามจำนวนคอลัมภ์ของข้อมูลชุดฝึกสอน

การคูณค่าน้ำหนักค่าน้ำหนักสุ่มคอลัมภ์ที่ 1 กับทุกแถวข้อมูลอินพุตคอลัมภ์ที่ 1

การคูณค่าน้ำหนักค่าน้ำหนักสุ่มคอลัมภ์ที่ 2 กับทุกแถวข้อมูลอินพุตคอลัมภ์ที่ 2

การคูณค่าน้ำหนักค่าน้ำหนักสุ่มคอลัมภ์ที่ 3 กับทุกแถวข้อมูลอินพุตคอลัมภ์ที่ 3

การคูณค่าน้ำหนักค่าน้ำหนักสุ่มคอลัมภ์ที่ 4 กับทุกแถวข้อมูลอินพุตคอลัมภ์ที่ 4

การคูณค่าน้ำหนักค่าน้ำหนักสุ่มคอลัมภ์ที่ 5 กับทุกแถวข้อมูลอินพุตคอลัมภ์ที่ 5

ภาพประกอบที่ 5-24 ตัวอย่างการสุ่มค่าน้ำหนักในการทดสอบประเภทที่ 2

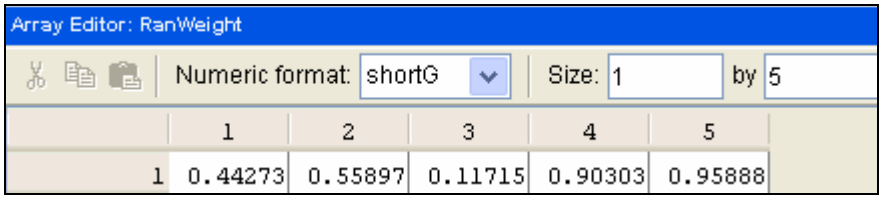
การสุ่มค่าน้ำหนักที่จะนำไปคูณกับข้อมูลอินพุต เพื่อสร้างเป็นค่าน้ำหนักเริ่มต้นใหม่ นั้น จะเป็นการอาศัยคำสั่ง rand ใน Matlab ดังภาพประกอบ 5-24 ในการสุ่มค่าตัวเลขที่อยู่ในช่วง 0-1 จำนวน 5 ค่า และค่าที่ได้จะอยู่ในรูปแบบของเลขทศนิยมดังภาพประกอบ 5-25

ฟังก์ชัน rand (มนัส สังวรศิลป,2542) จะทำการสร้างอะเรย์ของการสุ่มจำนวน (random numbers uniformly distributed) โดยผลลัพธ์ที่ได้จะเป็นข้อมูลเมตริกซ์ของตัวเลข ที่เป็นเมตริกซ์ขนาด [1x5] ที่แต่ละอีลิเมนต์เป็นค่าสุ่มจากตัวเลขในช่วง 0-1 โดย 1 คือจำนวนแถว และ 5 คือ จำนวนคอลัมภ์ และในงานวิจัยนี้จะเลือกใช้คุณสมบัติฟังก์ชันนี้ในการสุ่มค่าตัวเลข เพื่อนำไปคูณกับข้อมูลค่าน้ำหนักอินพุต เพื่อประเมินระดับค่าน้ำหนักความสูงต่ำของแต่ละค่าน้ำหนักพารามิเตอร์ในการเรียนรู้และทดสอบของโครงข่าย

การทดสอบการเรียนรู้และทดสอบของโครงข่ายผ่านโปรแกรม MatLab ในการทดสอบประเภทที่ 2 นี้ จะเป็นการให้โครงข่ายทำการเรียนรู้ซ้ำ ๆ และทำการปรับเปลี่ยนสุ่มค่าตัวเลขไปเรื่อย ๆ เป็นจำนวน 1000 ครั้ง โดยในแต่ละครั้งนั้นจะให้โครงข่ายเรียนรู้จำนวน 1000 รอบต่อครั้ง เพื่อการหาค่าน้ำหนักสุ่มที่เหมาะสมที่เมื่อนำมาคูณกับค่าน้ำหนักอินพุตและทำให้ผลในการทำนายมีค่าความถูกต้องสูงที่สุด โดยจะทำการคัดเลือกเฉพาะค่า น้ำหนักที่มีค่าความถูกต้องในการสอน และค่าความถูกต้องในการทดสอบ ค่อนข้างดีหรือมีค่าสูง โดยมีขั้นตอนดังนี้

4.8.1.2.1 ขั้นตอนการสุ่มค่าน้ำหนัก

การสุ่มค่าน้ำหนักโดยทำการสุ่มค่าน้ำหนักขึ้นมาในรูปแบบของเมตริกซ์ [1 x 5] โดยจะทำการสุ่มค่าน้ำหนักและผลที่ได้จากการสุ่ม คือ “ ค่าน้ำหนักสุ่ม = [ค่าสุ่ม1 ค่าสุ่ม2 ค่าสุ่ม3 ค่าสุ่ม4 ค่าสุ่ม5] ” ดังภาพประกอบ 5-25



ภาพประกอบที่ 5-25 ตัวอย่างค่าตัวเลขที่ได้จากการสุ่ม

4.8.1.2.12 ขั้นตอนการคูณค่าน้ำหนัก

ขั้นตอนการคูณค่าน้ำหนักนั้น เป็นการนำน้ำหนักที่สุ่มมาคูณกับข้อมูลค่าน้ำหนักของคำหรือวลีในข้อมูลชุดฝึกสอนที่ป้อนเข้าสู่โครงข่าย เพื่อให้เกิดการปรับเปลี่ยนค่าน้ำหนักของข้อมูลลชุดสอน ที่สามารถทำให้โครงข่ายเรียนรู้ได้ดีที่สุดและส่งผลให้การสกัดคำหรือวลีสำคัญมีความถูกต้องยิ่งขึ้น โดยจะไม่คูณค่าน้ำหนักสุ่มกับข้อมูลชุดทดสอบ เพราะข้อมูลชุดทดสอบนั้นจะต้องเป็นค่าคงที่เสมอเพื่อให้โครงได้ทำการทดสอบและทำการประเมินค่าที่ใกล้เคียงที่สุด และเป็นข้อมูลที่โครงข่ายไม่จำเป็นจะต้องใช้ในการเรียนรู้ แต่จะเป็นข้อมูลที่โครงข่ายใช้สำหรับทำนาย โดยมีการจับคู่การคูณค่าน้ำหนักในข้อมูลชุดฝึกสอนดังนี้

- ข้อมูลค่าน้ำหนักความถี่ คูณด้วยค่าน้ำหนักสเกล่าร์สุ่มตำแหน่งที่ 1
- ข้อมูลค่าน้ำหนักตำแหน่งในประโยค คูณด้วยค่าน้ำหนักสเกล่าร์สุ่มตำแหน่งที่ 2
- ข้อมูลค่าน้ำหนักตำแหน่งในย่อหน้า คูณด้วยค่าน้ำหนักสเกล่าร์สุ่มตำแหน่งที่ 3
- ข้อมูลค่าน้ำหนักตำแหน่งในเอกสาร คูณด้วยค่าน้ำหนักสเกล่าร์สุ่มตำแหน่งที่ 4
- ข้อมูลค่าน้ำหนักหน้าที่คำ คูณด้วย ค่าน้ำหนัก สเกล่าร์ สุ่ม ตำแหน่งที่ 5

4.8.1.2.3 การนำค่าน้ำหนักเริ่มต้นไปทดสอบ

ขั้นตอนสุดท้ายของการทดสอบประเภทที่ 2 นี้ คือ การนำน้ำหนักเริ่มต้นใหม่ ที่ได้จากการคูณด้วยค่าน้ำหนักสุ่มแล้วนั้น สำหรับเป็นค่าน้ำหนักในข้อมูลชุดฝึกสอนให้กับโครงข่าย (ชุดใหม่) โดยหลังจากทำการทดสอบจนครบจำนวนครั้งที่กำหนดไว้แล้วนั้น จึงจะนำผลที่ได้มาทำการวิเคราะห์ถึงค่าน้ำหนักที่เหมาะสมและดีที่สุดในการสกัดคำหรือวลีสำคัญที่เกิดขึ้นในแต่ละชุดข้อมูลการทดสอบ

4.9 การประเมินผลคำสำคัญ

โดยทำการประเมินผล ค่าความแม่นยำในการระบุคำหรือวลีสำคัญของเอกสารชุดทดสอบ ดังนี้

$$\text{ค่าความถูกต้องของข้อมูล (\%)} = \frac{\text{จำนวนแถวข้อมูลที่ทำนายถูก}}{\text{จำนวนแถวข้อมูลทั้งหมด}} * 100$$