📖 03 - Contextual Text Analysis

ผศ. ดร. ชนันท์กรณ์ จันแดง

สำนักวิชาสารสนเทศศาสตร์ มหาวิทยาลัยวลัยลักษณ์

🧭 บทนำ: การวิเคราะห์ข้อความเชิงบริบท

แนวคิดหลัก:

เข้าใจ "ความหมายของข้อความ" โดยคำนึงถึงบริบท (Context)

> คำว่า “bank” อาจหมายถึง “ธนาคาร” หรือ “ริมฝั่งน้ำ”
> บริบทจึงมีความสำคัญต่อการตีความ

⚙ Text Embedding คืออะไร

Text Embedding: การแปลงข้อความให้อยู่ในรูปเวกเตอร์มิติสูง
ข้อความที่มีความหมายคล้ายกัน → เวกเตอร์อยู่ใกล้กัน

- Cosine Similarity → วัดมุมระหว่างเวกเตอร์
- Euclidean Distance → วัดระยะทางระหว่างเวกเตอร์

> “ครู” ใกล้กับ “อาจารย์” และ “ผู้สอน”

🧩 ทฤษฎีพื้นฐานทางวิชาการ

Distributional Semantics Theory

> “You shall know a word by the company it keeps.”
> (Firth, 1957)

- คำที่อยู่ในบริบทเดียวกันมักมีความหมายคล้ายกัน
- เป็นรากฐานของแนวคิดการเรียนรู้เชิงการกระจาย

Vector Space Model (VSM)

- การแทนข้อความในรูปเวกเตอร์
- วัดความคล้ายคลึงด้วย cosine / distance
- เป็นจุดเริ่มต้นของแนวคิด Embedding

Word Embedding Models

โมเดล	แนวคิด	จุดเด่น
Word2Vec	เรียนรู้คำจากบริบท	เข้าใจความสัมพันธ์เชิงความหมาย
GloVe	ใช้ Co-occurrence Statistics	เน้น global context
FastText	ใช้ subword	รองรับคำใหม่ (OOV words)

Contextual Embeddings

- ใช้ Deep Neural Networks (LSTM / Transformer)
- คำเดียวกันมีเวกเตอร์ต่างกันในบริบทต่างกัน

โมเดล	ลักษณะ	จุดเด่น
ELMo	Bi-LSTM	บริบทสองทิศทาง
BERT	Transformer สองทิศทาง	เข้าใจจากซ้าย-ขวา
GPT	Transformer เดียว	สร้างข้อความต่อเนื่อง
SBERT	ปรับ BERT เพื่อเปรียบเทียบประโยค	ใช้ใน Semantic Search

🔬 แนวคิดทฤษฎีขั้นสูง

Manifold Hypothesis

- ข้อความมีโครงสร้างซับซ้อนในมิติสูง
- Embedding คือการเรียนรู้โครงสร้างภายใน (semantic manifold)

Transfer Learning

- ใช้ embedding ที่ฝึกมาก่อน เช่น BERT, GloVe
- นำไปใช้ fine-tuning ในงานใหม่ (เช่น Sentiment Analysis)

Explainability

- พยายามตีความแกนของเวกเตอร์ เช่น เพศ, อารมณ์

💼 การประยุกต์ใช้ Text Embedding

การประยุกต์	คำอธิบาย
Semantic Search	ค้นหาตามความหมายแทนการค้นหาคำตรงตัว
Clustering	จัดกลุ่มข้อความใกล้เคียง
Recommendation	แนะนำเนื้อหาคล้ายกัน
Similarity Measurement	ตรวจจับการลอกงาน / ความเหมือนของประโยค
LLM Understanding	ใช้ใน GPT, BERT, Claude ฯลฯ

🧠 สรุปเชิงทฤษฎี

1. Embedding = สะพานระหว่างภาษาและคณิตศาสตร์
2. Contextual Embedding = วิวัฒนาการขั้นสูงของ NLP
3. เป็นรากฐานของระบบ Search, Chatbot, และ LLMs

🧪 Workshop: Text Embedding

วัตถุประสงค์

- เข้าใจการแปลงข้อความเป็นเวกเตอร์
- วัดความคล้ายคลึงเชิงความหมาย
- เปรียบเทียบ Keyword vs Contextual Approach

🔧 เครื่องมือและข้อมูล

- Python / Google Colab
- Libraries: `sentence-transformers`, `numpy`, `sklearn`

```
texts = [
 "รถยนต์ไฟฟ้าเป็นมิตรต่อสิ่งแวดล้อม",
 "ยานยนต์พลังงานสะอาดกำลังได้รับความนิยม",
 "ธนาคารให้บริการสินเชื่อบ้าน",
 "แม่น้ำมีน้ำมากในฤดูฝน"
]
```

🧠 สร้าง Sentence Embeddings

```
from sentence_transformers import SentenceTransformer
model = SentenceTransformer('paraphrase-multilingual-MiniLM-L12-v2')
embeddings = model.encode(texts)
```

- ข้อความ → เวกเตอร์มิติ 384
- ความหมายใกล้กัน → เวกเตอร์อยู่ใกล้กัน

📏 Semantic Similarity

```
from sentence_transformers import util
similarity = util.cos_sim(embeddings, embeddings)
```

ค่า	การตีความ
0.8–1.0	คล้ายกันมาก
0.5–0.8	ค่อนข้างคล้าย
<0.2	ต่างกันมาก

📊 Visualization

```
from sklearn.manifold import TSNE
plt.scatter(...)
```

- จุดใกล้กัน → ข้อความคล้ายกัน
- แสดง semantic clusters

🧪 Showcase: Spam Detection

แนวคิด:

แยกข้อความเป็น 2 กลุ่ม — Spam vs Ham

ใช้ Contextual Embedding + KNN Classifier

```
data = {...}
model = SentenceTransformer('all-MiniLM-L6-v2')
embeddings = model.encode(data['text'])
```

🔍 ตัวอย่างผลลัพธ์

ข้อความ	ผลลัพธ์
“Get rich quick!”	spam
“Review project report.”	ham

Embedding เข้าใจเจตนา (intent) แม้ไม่มี keyword “spam”

🧠 Key Insights

- Context สำคัญกว่าคำเดี่ยว
- Embedding ช่วยให้คอมพิวเตอร์เข้าใจ “ความหมาย”
- ใช้ได้ทั้งในงาน Search, Chatbot, และ AI Model
- เป็นรากฐานของ LLM ทุกชนิด

🙏 ขอบคุณครับ

Assist. Prof. Dr. Chanankorn Jandaeng

Walailak University

Contextual Text Analysis & Embedding Workshop