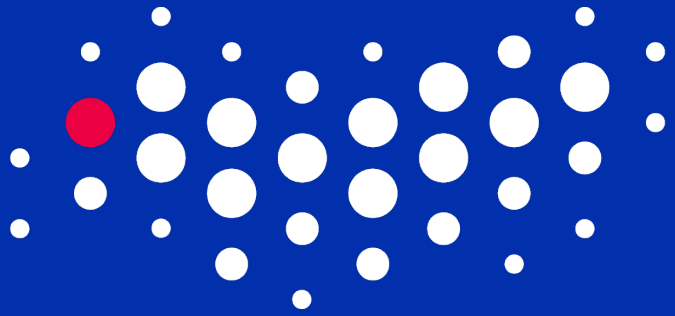

ITMO UNIVERSITY

NLP Basic and Selected Topics

A Practical and Easy Introduction to Selected Topics

Aug 31st, 2019

Overview of the Unit Today

1) Applications of NLP / Introduction (30min)

2) Practical NLP (NLTK / pythainlp) (45min)

3) Modern NLP with ML/DL (45min)

4) Example: Word Similarity and WordNet (30min)

5) Modern NLP with fastAI / flair (30min)

Word Similarity

- What does it mean?
  - How similar two words are.
- Why is it important?
  - If somebody can understand how similar two words are, then own can understand some basics of a language
  - If I system understands how similar two words are, then it has some understanding of a language

Word Similarity

- Applications of word similarity
  - For example in **search**
- Word embedding models can be evaluated by testing how good the similarity in the model corresponds to real similarity
- **How can we define real similarity – what would you do to test a system?**

Word Similarity Datasets

- As real “gold standard” similarity we can create datasets
- The datasets contain 2 words, and a score (for example from 1-10) how similar they are

- **How would you set the score using human assessment?**

Word Similarity Datasets

- **How would you set the score using human assessment?**
- For example 10 people can rate, and then we take the average

WordSim-353 example

Word 1	Word 2	Human (mean)
tiger	cat	7.35
tiger	tiger	10.00
sugar	approach	0.88
book	paper	7.46
stock	egg	1.81
computer	keyboard	7.62
computer	internet	7.58
plane	car	5.77
train	car	6.31

Word Similarity Datasets

- There are many datasets for English: WordSim-353, SimLex-999, SemEval-500 ...
- The datasets have slightly different characteristics, and different scales ..

Thai language datasets

- This year we (Netiposakul, Wohlgenannt) translated 3 English datasets to Thai
- After translation, new ratings of similarity
- The new datasets: **TH-WordSim-353, TH-SimLex-999, TH-SemEval-500**
- https://arxiv.org/abs/1904.04307

Thai language datasets

- https://github.com/gwohlgen/thai_word_similarity

```
เก่า,ใหม่,2.19
หลักแหลม,ฉลาด,8.44
ยาก,ยาก,10
สุข,ร่าเริง,6.67
ยาก,ง่าย,2.29
ด่วน,รวดเร็ว,7.19
```

Thai language datasets

- We evaluated different Thai word embedding models with this datasets
- Another option is to use structured sources like WordNet to compute similarity scores

Question

- **How can we measure the quality of the model (word embedding or WordNet) with regards to dataset??**

Correlation

- What is it?
- Give examples from different domains:
  - Sport
  - Medicine
- How to measure it?
- What scale (interval) used?

Correlation

- Give examples from different domains:
  - For example: IQ / income
  - Medicine: weight / diabetes
- How to measure it? Pearson / Spearman
- What scale (interval) used? [-1 ... 1]

WordNet

- What is WordNet?
  - A lexical database that connect word and their meanings
- In WordNet, a word can have a number of meanings, the meanings are called **synsets**
- Then those synsets are connected by a number of relations, like hypernymy, antonymy, etc.

WordNet

- WordNet is also integrated into Python
- **Show WordNet basic functions (see below)**
- http://www.nltk.org/howto/wordnet.html

Exercise

รับ,ให้,3.02
แนะนำ,แนะนำ,10
เลียนแบบ,วาดภาพ,1.88
คิด,ตัดสิน,2.71
ทักทาย,พบ,2.29

- Start from this part of the dataset
- Compute path_similarity in WordNet (use first synset for word
- Save results
- Compute Pearson https://kite.com/python/examples/656/scipy-compute-the-pearson-correlation-coefficient