NATURAL LANGUAGE ANALYSIS

LESSON 6: SIMPLE SEMANTIC ANALYSIS

OUTLINE

- What is Semantic?
- Content Analysis
- Semantic Analysis in CENG
- Semantic Analysis in NLP
- Vector Space Model
- Semantic Relations
- Latent Semantic Analysis (LSA)

WHAT IS SEMANTIC?

- Semantic is the meaning, interpretation of the words, signs and sentence structure.
- As you see in the figure, saying hello is different according to languages but meaning is the same.
- So semantic deals with the meaning of the things that is saved its behind.

WHAT IS SEMANTIC?

There are two types of meaning in a language. They are conceptual meaning and associative meaning.

- Semantic deals with conceptual meaning. This is also known as dictionary definition of the concept.
- Associative meaning is also known as Pragmatic and interest in the study of how context affects meaning.
- For conceptual meaning, **needle** means ‘**thin, sharp, steel instrument**’. But in associative meaning, needle =‘painful’.

CONTENT ANALYSIS

- Content analysis is a formal methodology to study a collection of media to discover, uncover, or answer
- Content analysis can be carried out
  - Quantitatively
  - Qualitatively.

QUANTITATIVE ANALYSIS

- Counting and statistics: Numeric measurements
- Word frequencies: how many times does a word appear?
- Specify stop-words to ignore (e.g., the, and, others)
- Need to consolidate synonyms, stems (e.g., dog = dogs)
- Compound words (i.e., word pairs) are important
  - United States
  - not good

QUALITATIVE ANALYSIS

- Coding is performed to reduce text collection to categories (i.e., concepts)
- Analyst can seed concepts or discover concepts during analysis
- Often, the more discovery allowed the more objective the analysis (grounded theory reduces researcher bias)
- Concepts and their relationships form the foundations for extracting meaning

SEMANTIC ANALYSIS IN CENG

There are lexical analysis, syntax analysis and semantic analysis phases in compiler design.

- Lexical analysis-> check the lexicons in the language, detects illegal inputs
- Syntax analysis-> using regular expressions of the language, check the syntax of each line in language, like variable definition, assignments, mathematical operations etc.;
- Semantic analysis-> it is the last, catching all errors before going into machine level like below;
- Checking variable types while assign a value to a variable;

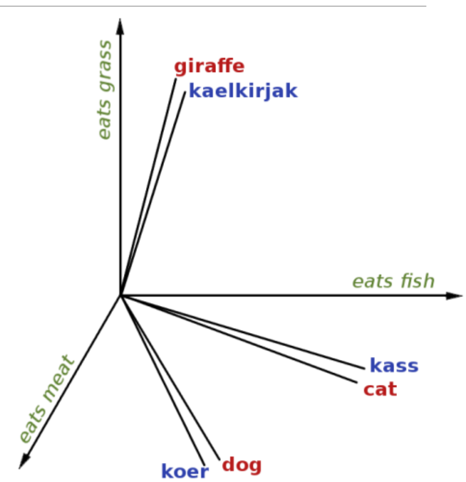
- string a;
- int b;
- b=a;
- This is also semantic analysis issue…
- --
- string[] a=new string[30];
- a[35]='asdf';
- This is also semantic analysis issue

SEMANTIC ANALYSIS IN NLP

- Semantic analysis of the word level is generally done for the word sense disambiguation, semantic similarity/relatedness.
- Sentence and short text analysis is generally done to get similarity (relatedness) of two given textual items, sentiment analysis, named entity recognition.
- Semantic analysis of the documents are generally done to get document similarity or relatedness, document classification, textual entailment, information retrieval, information extraction etc.

VECTOR SPACE MODEL

Vector Space Model represents each document, text, sentence, or word by a high-dimensional vector in the space of words



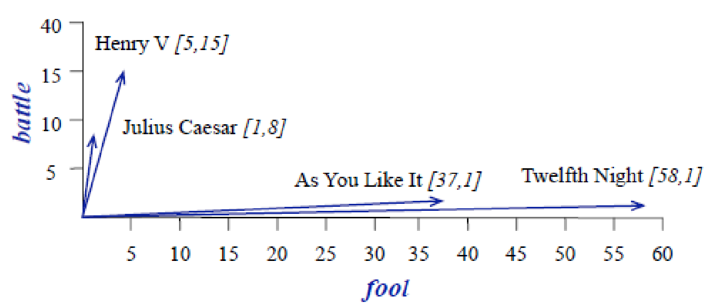
VECTOR SPACE MODEL

- The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.
- We can think of the vector for a document as identifying a point in |Vector|-dimensional space; thus the documents in table above are points in 4-dimensional space.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

VECTOR SPACE MODEL

- Since 4-dimensional spaces are hard to display here,
- Shows a visualization in two dimensions; we've arbitrarily chosen the dimensions corresponding to the words ***battle*** and ***fool***.



WORD VECTORS

- Documents can also be represented as vectors in a vector space.
- Vector semantics can also be used to represent the meaning of words, by associating each word with a vector.
- The word vector is now a row vector rather than a column vector and hence the dimensions of the vector are different.
- The four dimensions of the vector for **fool**, [37,58,1,5], correspond to the four Shakespeare plays.

WORD VECTORS

- Each entry in the vector thus represents the counts of the word's occurrence in the document corresponding to that dimension.
- For documents, we saw that similar documents had similar vectors, because similar documents tend to have similar words.
- This same principle applies to words: similar words have similar vectors because they tend to occur in similar documents.
- The term-document matrix thus lets us represent the meaning of a word by the documents it tends to occur in.

WORD TO WORD MATRIX OR TERM-CONTEXT MATRIX

- The context could be the document, in which case the cell represents the number of times the two words appear in the same document.
- It is most common, however, to use smaller contexts, generally a window around the word, for example of 4 words to the left and 4 words to the right,
- Below slide a figure represents the number of times (in some training corpus) the column word occurs in such a ±4 word window around the row word.

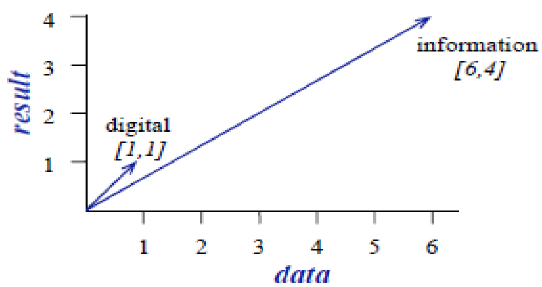
WORD TO WORD MATRIX OR TERM-CONTEXT MATRIX

- Co-occurrence vectors for four words, computed from the Brown corpus, showing only six of the dimensions. The vector for the word **digital** is outlined in red. Note that a real vector would have vastly more dimensions and thus be sparser.

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

WORD TO WORD MATRIX OR TERM-CONTEXT MATRIX

A spatial visualization of word vectors for **digital** and **information**, showing just two of the dimensions, corresponding to the words **data** and **result**.



WORD TO WORD MATRIX OR TERM-CONTEXT MATRIX

- Note that |V|, the length of the vector, is generally the size of the vocabulary, usually between 10,000 and 50,000 words.
- But of course since most of these numbers are zero these are sparse vector representations, and there are efficient algorithms for storing and computing with sparse matrices.
- The size of the window used to collect counts can vary based on the goals of the representation, but is generally between 1 and 8 words on each side of the target word (for a total context of 3-17 words).
- In general, the shorter the window, the more syntactic the representations, since the information is coming from immediately nearby words; the longer the window, the more semantic the relations.

WEIGHTING TERMS

- While representing document vectors or word vectors, terms in the documents are weighted or normalized.
- One of the main methods for term weighting is the TF-IDF.
- Mostly, terms in the documents are normalized between [0 1].

MEASURING SEMANTIC SIMILARITY

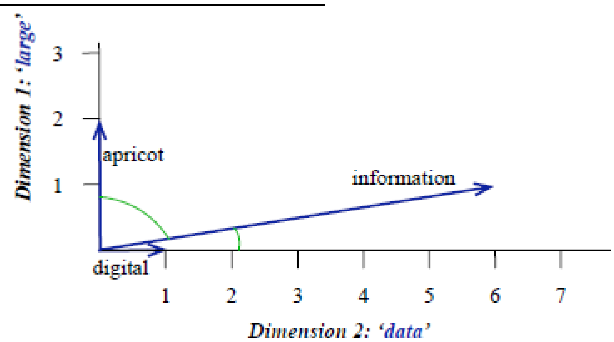
- To define similarity between two target words v and w, we need a measure for taking two such vectors and giving a measure of vector similarity.
- By far the most common similarity metric is the cosine of the angle between the vectors.

MEASURING SEMANTIC SIMILARITY

	large	data	computer
apricot	2	0	0
digital	0	1	2
information	1	6	1

$$\cos(\text{apricot}, \text{information}) = \frac{2+0+0}{\sqrt{4+0+0}\sqrt{1+36+1}} = \frac{2}{2\sqrt{38}} = .16$$

$$\cos(\text{digital}, \text{information}) = \frac{0+6+2}{\sqrt{0+1+4}\sqrt{1+36+1}} = \frac{8}{\sqrt{38}\sqrt{5}} = .58$$



SEMANTIC RELATIONS

- Semantic relationships are the associations that there exist between the meanings of words (semantic relationships at word level), between the meanings of phrases, or between the meanings of sentences (semantic relationships at phrase or sentence level).

Relationship Type	Example
Equivalency	
Synonymy	UN / United Nations
Lexical variants	pediatrics / paediatrics
Near synonymy	sea water / salt water smoothness / roughness
Hierarchy	
Generic or IsA	birds / parrots
Instance or IsA	sea / Mediterranean Sea
Whole / Part	brain / brain stem
Associative	
Cause / Effect	accident / injury
Process / Agent	velocity measurement / speedometer
Process / Counter-agent	fire / flame retardant
Action / Product	writing / publication
Action / Property	communication / communication skills
Action / Target	teaching / student
Concept or Object / Property	steel alloy / corrosion resistance
Concept or Object/ Origins	water / well
Concept or Object / Measurement Unit or Mechanism	chronometer / minute
Raw material / Product	grapes / wine
Discipline or Field / Object or Practitioner	neonatology / infant

SEMANTIC CLASSIFICATION

- In order to classify the documents, basic method is the comparison of the document words with the given keyword list of the each topics.
- Maximum number of keywords from a topic may determine the topic of the documents.

SEMANTIC CLASSIFICATION

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

LATENT SEMANTIC ANALYSIS (LSA)

- LSA is a famous text classification method.
- LSA aims to discover something about the meaning behind the words; about the topics in the documents.
- What is the difference between topics and words?
  - Words are observable
  - Topics are not. They are latent.
- How to find out topics from the words in an automatic way?
  - We can imagine them as a compression of words
  - A combination of words

LATENT SEMANTIC ANALYSIS (LSA)

- Uses Singular Value Decomposition (SVD) to simulate human learning of word and passage meaning.
- Represents word and passage meaning as high-dimensional vectors in the semantic space.
- Implements the idea that the meaning of a passage is the sum of the meanings of its words.
- meaning of word_1 + meaning of word_2 + … + meaning of word_n = meaning of passage
- By creating an equation of this kind for every passage of language that a learner observes, we get a large system of linear equations.

HOW LSA WORK

- Takes as input a corpus of natural language
- The corpus is parsed into meaningful passages (such as paragraphs)
- A matrix is formed with passages as rows and words as columns. Cells contain the number of times that a given word is used in a given passage.
- The cell values are transformed into a measure of the information about the passage identity the carry

HOW LSA WORK

	d1	d2	d3	d4	d5	d6
cosmonaut	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

SINGULAR VALUE DECOMPOSITION

- SVD is applied to re-represent the words and passages as vectors in a high dimensional space.
- Real data usually have thousands, or millions of dimensions
  - E.g., web documents, where the dimensionality is the vocabulary of words
  - Facebook graph, where the dimensionality is the number of users.
- Huge number of dimensions causes problems
- The complexity of several algorithms depends on the dimensionality and they become infeasible.

SINGULAR VALUE DECOMPOSITION

$$\underset{[n\times m]}{A} = \underset{[n\times r]}{U}\ \underset{[r\times r]}{\Sigma}\ \underset{[r\times m]}{V^T} = [u_1, u_2, \cdots, u_r]\begin{bmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_r \end{bmatrix}\begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{bmatrix}$$

r: rank of matrix A

$\sigma_1, \geq \sigma_2 \geq \cdots \geq \sigma_r$: singular values of matrix A (also, the square roots of eigenvalues of AA^T and A^TA)

$u_1, u_2, \dots, u_r$: left singular vectors of A (also eigenvectors of AA^T)

$v_1, v_2, \dots, v_r$: right singular vectors of A (also, eigenvectors of A^TA)

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$$