UNIVERSITY of
HOUSTON
DEPARTMENT OF COMPUTER SCIENCE

NLP for Security Module

Rakesh Verma

Introduction to Natural Language Processing

A word about text mining

Bag of words model of text

Preprocessing may include: stemming, stopword elimination, lowercasing, etc.

Vector space model of sentences, documents.

TF-IDF popular method, should be used as baseline for most tasks

Historical Tidbit

Panini, Sanskrit philologist, grammarian, composes Ashtadhyayi around 4th-6th century BCE.

Considered “father of linguistics”

3959 verses or rules on linguistics, syntax and semantics in 8 chapters

17th century manuscript of Panini’s grammar

[Source: Wikipedia]

Why study Natural Language Processing (NLP)?

NLP techniques are being used in:

Email Classification

Opinion mining and opinion fraud detection

Determination of strong passwords

…

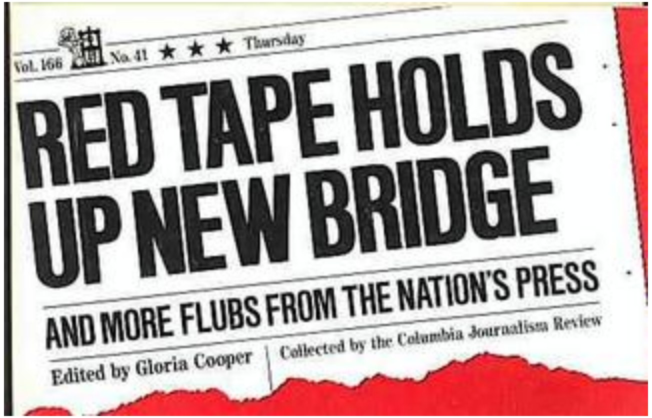
NLP-based Detection

Natural Language Processing is challenging

Natural languages have polysemous words, collocations, ambiguity, exceptions, …

One needs a little knowledge of the world to understand the meaning of a single sentence

.



Vol. 166 No. 41 ★ ★ ★ Thursday

RED TAPE HOLDS UP NEW BRIDGE

AND MORE FLUBS FROM THE NATION'S PRESS

Edited by Gloria Cooper | Collected by the Columbia Journalism Review

A Translation Experiment

Input	Output
Computers are not good at crunching text	**컴퓨터가 텍스트를 재정에서 잘되지 않습니다**
컴퓨터가 텍스트를 재정에서 잘되지 않습니다	Текст, компьютер не знает финансовый
Текст, компьютер не знает финансовый	உரை, கணினி நிதி தெரியாது
உரை, கணினி நிதி தெரியாது	Text, the system does not fund

With Chinese replacing Korean in the experiment we get?

The text is not very good mathematical systems - 2012

On 4/13/2017 at 4.10am IST in Chennai

Input	Output
Computers are not good at crunching text	**컴퓨터가 텍스트를 잘 처리하지 못함**
컴퓨터가 텍스트를 잘 처리하지 못함	Ошибка компьютера совладать текст
Ошибка компьютера совладать текст	கணினிப் பூச்சியைக் கட்டுப்பாடு உரை
கணினிப் பூச்சியைக் கட்டுப்பாடு உரை	Computer bug control speech

What is Natural Language Processing (NLP)?

- Natural Language Processing is a subfield of computer science, artificial intelligence, and computational linguistics that focuses on language processing.
  - Language being both human and computer languages
- Major applications of NLP
  - Automatic Summarization
  - Machine Translation
  - Sentiment Analysis
  - Speech Recognition
- Notable apps that utilize NLP methods
  - Siri, Cortana, Okay Google, Amazon Echo
  - Dragon Naturally Speaking
  - Ask.com (Ask Jeeves)

Linguistic Objectives

- What kinds of things do people say?
  - What is the structure of language?
  - Usage patterns of words can carry deep meaning if they can be leveraged.
- What do these things say about the world?
  - What do the words we say relate to in reality?
  - What connections do they have to meaning, context, and communication?

Linguistic Objectives (Continued)

Traditional linguistics tries to answer these questions on the basis of underlying **competence grammars**.

A grammar is the set of rules that govern the structure of sentences in the language.

The **grammaticality** of sentences is whether or not the sentence is recognized by the associated grammar. Other sentences are deemed **ungrammatical**.

A sentence can be grammatical and but not carry any clear meaning.

"Colorless green oranges run slowly."

The **conventionality** of sentences is whether or not they conform to common presentations of the same ideas.

Two grammatical sentences with the same meaning, but one of them sounds better.

Issues with this Categorical View

Grammars are rigid in theory, flexible in practice.

You can still grasp the inherent meaning of sentences when the grammar is incorrect.

Determining grammaticality is rather difficult for people to do for longer complex sentences.

"Those are the books you should read before it becomes difficult to talk about."

"That a serious discussion could arise here of this topic was quite unexpected."

Those are both grammatical sentences. If you disagree, that's the point.

This categorical view of language is not perfect, but is sufficient for many NLP purposes.

For less complex sentences, we should still be able to determine their grammaticality.

For more complex sentences, more robust methods may be required.

Language as Probabilistic Phenomena

Interactions with the world are always conducted with imperfect information.

Language is just another medium through which we as people take in information and incorporate it into our decision making.

- If someone says, "The benches in the park were just painted," we know not to sit on them even before we get to the park.

Comprehension of speech is therefore modelled nicely as probabilistic processes.

Unfortunately, grammaticality is not approachable probabilistically.By considering historical contexts, sentences that have never be said are equally unlikely to be said. Novel ungrammatical sentences are equally as likely as novel grammatical ones.

The Ambiguity of Language

NLP systems are commonly built to answer "**Who did what to whom?**"

Parts of speech describe the syntactic function of a word in a sentence.

For example, in the above sentence 'describe' is verb.

Some words have multiple possible parts of speech.

Example: bow your head (verb), a bow in your hair (noun)

Some sentences remain grammatical under different parts of speech.

"I saw her duck." is a sentence where 'duck' is either a verb or a noun.

Coordinating conjunctions without the serial comma.

In the sentence, "I'd like to thank my parents, the Academy and God." God could be interpreted as one of the speaker's parents, as could the Academy.

As sentences get longer, the number of ways to interpret them explodes.

Disambiguation

Selectional restrictions provide an avenue for disambiguating the true interpretation of a sentence.

We know that the verb "swallow" requires a physical object, so we can ignore interpretations where it isn't followed by one.

Unfortunately, this weakens the generalizability of our model to conventional metaphors.

"I swallowed his story, hook, line, and sinker."

"The supernova swallowed the planet."

Statistical NLP does not approach the problem with strict rules. Instead, it tries to encapsulate lexical and structural preferences from **corpora** using quantitative methods.

A **corpus** (pl. corpora) is a vast body of text, common terminology for an NLP dataset.

Lexical Resources

Lexical resources are any machine-readable text and any tools used to process them.

The *Brown corpus* is one such resource, compiled by Brown University in the 1960s and 1970s.

- The Brown corpus is a **balanced corpus** for American English.
- Balanced corpora are representative of their language across many contexts and genres.

The *Lancaster-Oslo-Bergen corpus* is the British English counterpart to the Brown corpus.

The *Susanne corpus* is a freely-available subset of the Brown corpus with the added benefit of annotated syntax structure information.

Lexical Resources (Continued)

The *Penn Treebank* is a corpus collected from the Wall Street Journal.

- It is a larger syntactically annotated corpus than the Susanne corpus, but it is not freely-available.

The *Canadian Hansards* is the best known example of a **bilingual corpus**.

- A bilingual corpus is one that contains **parallel texts**, texts are translations of each other in two or more languages.

WordNet is an freely-available electronic dictionary that organizes words into a hierarchy of **synsets**.

- Synsets are collections of words with identical or close meaning.
- It also carries additional relations between words based on potential meronymy.

Word Counts

When evaluating a text, you should calculate **word counts** and **word frequency**, the quantity and rate of appearance of individual words in text.

The most regularly appearing words are often **function words**, words that carry important grammatical information.

- Function words like determiners (ex. *mine*), prepositions (ex. *from*), and complementizers (ex. Mary believes *that* it is raining).

How many words are there in the text?

- **Word tokens** refer to the individual instances of each word in a text.
- **Word types** refer to the unique words used throughout.
- **Word frequencies** can be calculated by dividing by the total number of each respectively.

Word Counts (Continued)

In a text, not all words appear frequently enough.

- Word can appear as few times as once.
- These words are called **hapax legomana**, Greek for "read only once."

Though individual word types do not appear frequently on their own, all the rare words together can make up a significant portion of the text.

Word counts on rare words exemplify the difficulty of Statistical NLP.

- What can we learn about those words with so little data in the corpus?
- Will using a larger corpus solve this problem?

Unfortunately, this is not the case, as seen through **Zipf's Law**.

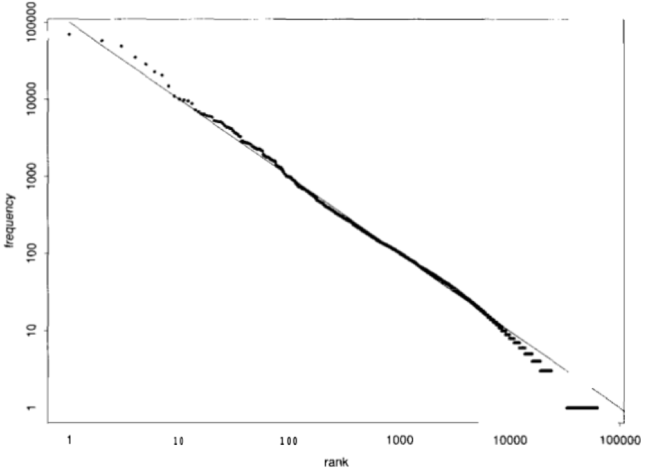
Zipf's Law

Zipf's Law is a famous law in linguistics that states that the there is a common "rough" ratio between word frequency and word rank.

Zipf's law is not as rigid as a physical law, but holds approximately *in general*.

$$frequency \propto \frac{1}{rank}$$

To the right, we have a graph of log-rank by log-frequency for the Brown Corpus.



As you can see, it's 'almost' linear, which indicates that rank and frequency are 'almost' proportional.

Zipf's Law (Continued)

Zipf's Law is based on the **Principle of Least Effort** which states that both the listener and the speaker are trying to minimize the work they do to communicate.

The speaker uses a small set of common words to construct comprehensible sentences easily.

The reader has a strong vocabulary so that each sentence can be unambiguously deciphered.

Collocations

A **collocation** is a turn of phrase that carries more meaning that each of its parts combined.

- Compounds (ex. *car wash*), phrasal verbs (ex. *make up*), and other common phrases (ex. *bacon and eggs*)

Collocations have an independent existence from their individual parts while they also show frequent ways in which a word is used.

A **bigram** is a pair of word tokens which appear in sequence with each other.

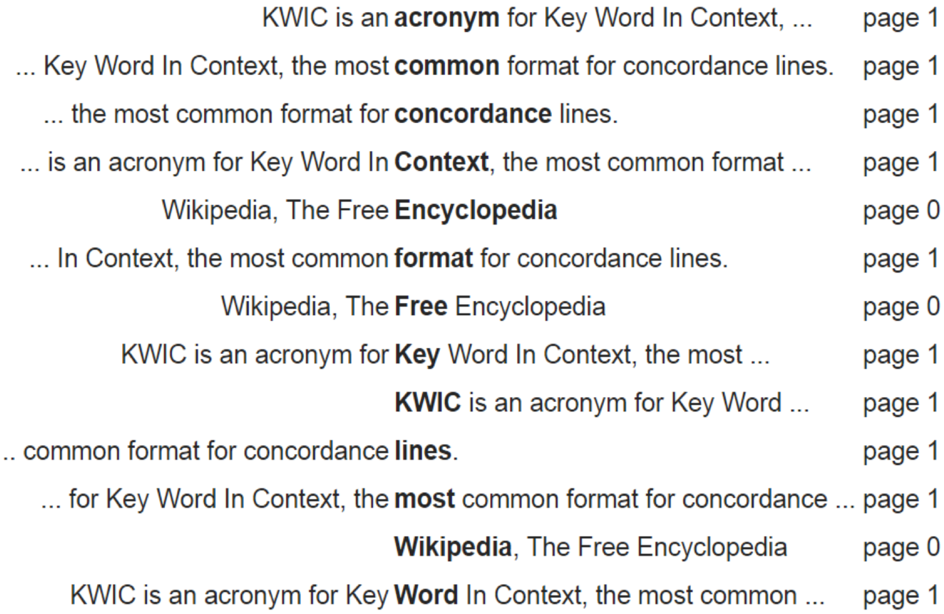
- Bigrams are not necessarily collocations, as many pairs are syntactic in nature (ex. "of the").
- A sequence of length *n* is called an **n-gram**.
  - Likewise, n-grams are not guaranteed to be a collocation, but collocations are n-grams for some *n*.

Concordances

A **concordance** is an alphabetical list of the principal words used in a body of work, listing every instance of each word with its immediate context.

Key Word in Context (KWIC) is the most common format, in which all words appear alphabetically.

To the right, is an example from the Wikipedia article for KWIC.



KWIC is an **acronym** for Key Word In Context, ...	page 1
... Key Word In Context, the most **common** format for concordance lines.	page 1
... the most common format for **concordance** lines.	page 1
... is an acronym for Key Word In **Context**, the most common format ...	page 1
Wikipedia, The Free **Encyclopedia**	page 0
... In Context, the most common **format** for concordance lines.	page 1
Wikipedia, The **Free** Encyclopedia	page 0
KWIC is an acronym for **Key** Word In Context, the most ...	page 1
KWIC is an acronym for Key Word ...	page 1
.. common format for concordance **lines**.	page 1
... for Key Word In Context, the **most** common format for concordance ...	page 1
Wikipedia, The Free Encyclopedia	page 0
KWIC is an acronym for Key **Word** In Context, the most common ...	page 1

References

Foundations of Statistical Natural Language Processing (Christopher D. Manning and Hinrich Schütze)