Welcome to:

Statistical Approaches

Unit objectives

After completing this unit, you should be able to:

- Understand what is statistical parsing and the core concepts involved in it
- Learn about multiword expressions and how to handle them
- Understand the concepts of word similarity and the relativeness calculations done
- Gain knowledge on word sense disambiguation and why it is needed in NLP
- Gain an insight into modern speech recognition techniques with an idea of the forerunners in the field
- Understand what statistical machine translation means and the guidelines needed to perform SMT

Parsing (1 of 2)

- An activity in computational linguistics and natural language processing identification and understanding of the syntax and semantics based on the grammar of the natural language.

- Parser is a tool for computation that can process any input sentence within the boundaries of the productions in the grammar and build structures called as parse trees within the confinement of the grammatical rule.

- Example: Sentence- Tom ate an apple.

```
sentence -> noun_phrase, verb_phrase
noun_phrase -> proper_noun
noun_phrase -> determiner, noun
verb_phrase -> verb, noun_phrase
proper_noun -> [Tom]
noun -> [apple]
verb -> [ate]
determiner -> [an]
```

Figure: Grammar

Source: https://forum.huawei.com/enterprise/en/what-is-parsing-in-nlp/thread/571685-100429

Parsing (2 of 2)

- Parsing of the statement constructs a parse tree with root, intermediate nodes, which are also called as non-terminal nodes, and leaf nodes called as terminal nodes.

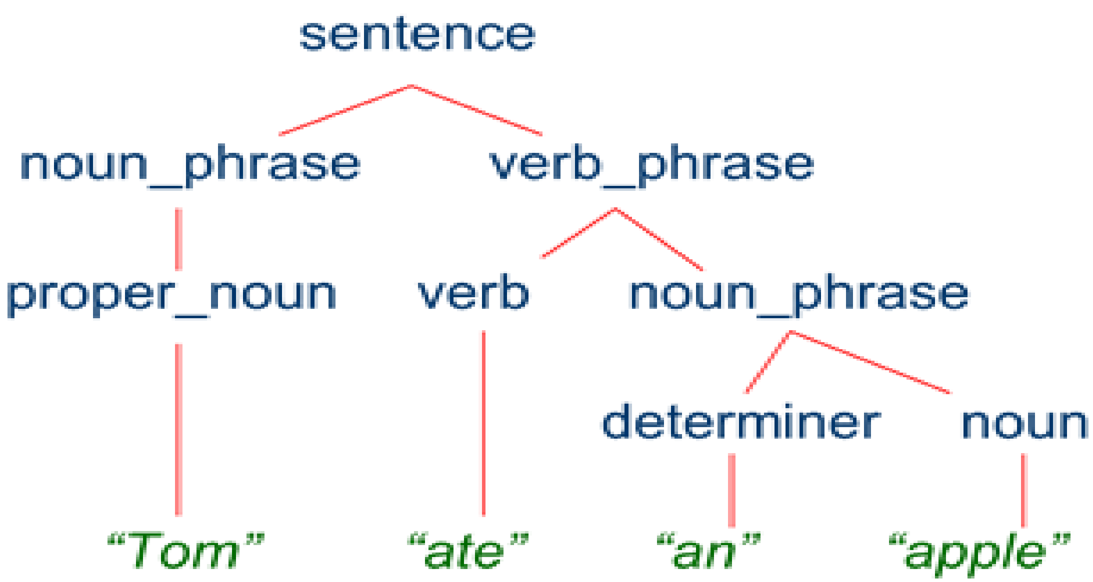


Figure: Parse Tree

Source: https://forum.huawei.com/enterprise/en/what-is-parsing-in-nlp/thread/571685-100429

Statistical parsing (1 of 2)

- Association between grammar of the natural language and the probability of its occurrence.
- Statistical parsing associates every grammar rule with a probability value.
- Example: Sentence: The can can hold water.

Statistical parsing (2 of 2)

- Large amount of grammatical rules ➔ Very large search space.

- Optimization on the subsets of the parse trees generated.

- Dissimilar parse trees ➔ Frequency identification.
- Similar parse trees ➔ Discarded.

- Statistical parsing ➔ Better performance ➔ Vocabulary is very large.

- Lexicalized and Statistical Parsing (LSP) ➔ Balance the vocabulary.

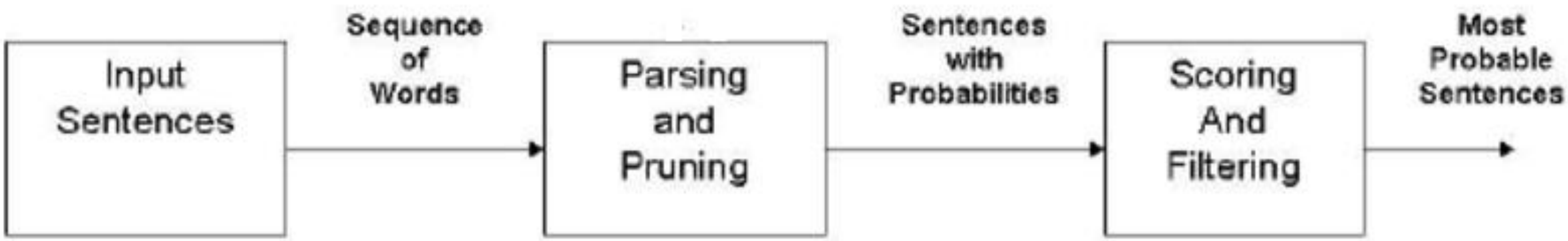


Figure: Statistical Parsing Steps

Source: https://www.researchgate.net/figure/Framework-of-Lexicalized-and-Statistical-Parser_fig1_220155897

Approaches to parsing

- Understands syntax and semantics.

- Parser ➔ Process an input sentence according to the production rules ➔ Parse trees.

Structural approach:

- Context free grammar (CFG) ➔ Group of consecutive words.

	Statistical	Structural
Foundation	Statistical decision theory	Human perception and cognition
Description	Quantitative features Fixed number of features Ignores feature relationships Semantics from feature position	Morphological primitives Variable number of primitives Captures primitive relationships Semantics from primitive encoding
Classification	Statistical classifiers	Parsing with syntactic grammars

Figure: Structural Approach vs Statistical Approach

Source: https://www.byclb.com/TR/Tutorials/neural_networks/ch1_1.htm

Statistical approach

- Statistical approaches are data driven.
- Concentrate upon short-term relationship between the words in a sentence.

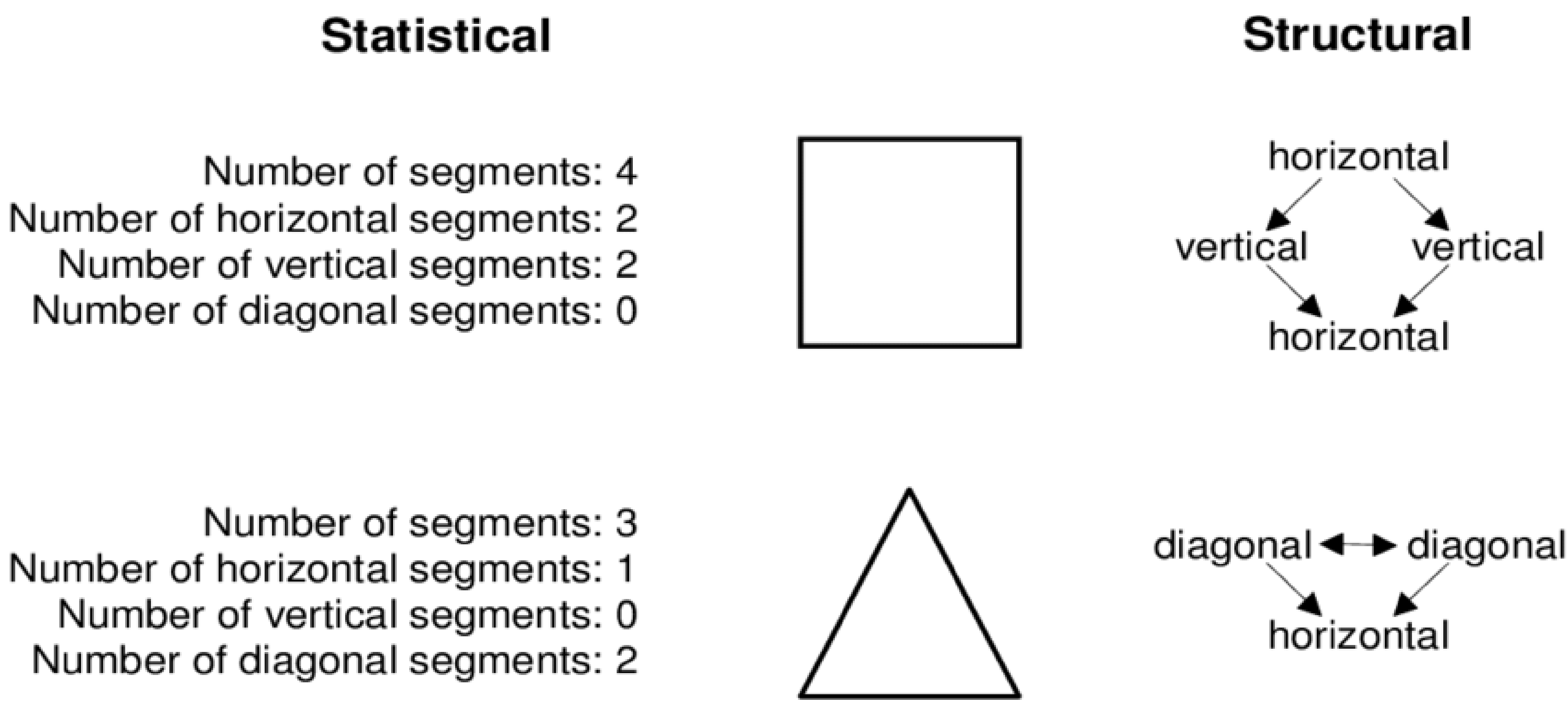


Figure: Analysis in Structural Approach vs Statistical Approach

Source: https://www.researchgate.net/figure/The-statistical-and-structural-approaches-to-pattern-recognition-applied-to-a-common_fig3_228558473

Lexicalized statistical parsing (1 of 2)

- Context free grammar is augmented using a probabilistic component and ambiguity is resolved in lexicalized statistical parsing.
- CFG is designed for adopting the probabilistic component into itself and is called as Probabilistic Context Free Grammar (PCFG).
- The performance of PCFG enhanced by adding a conditional rule for the lexical head.

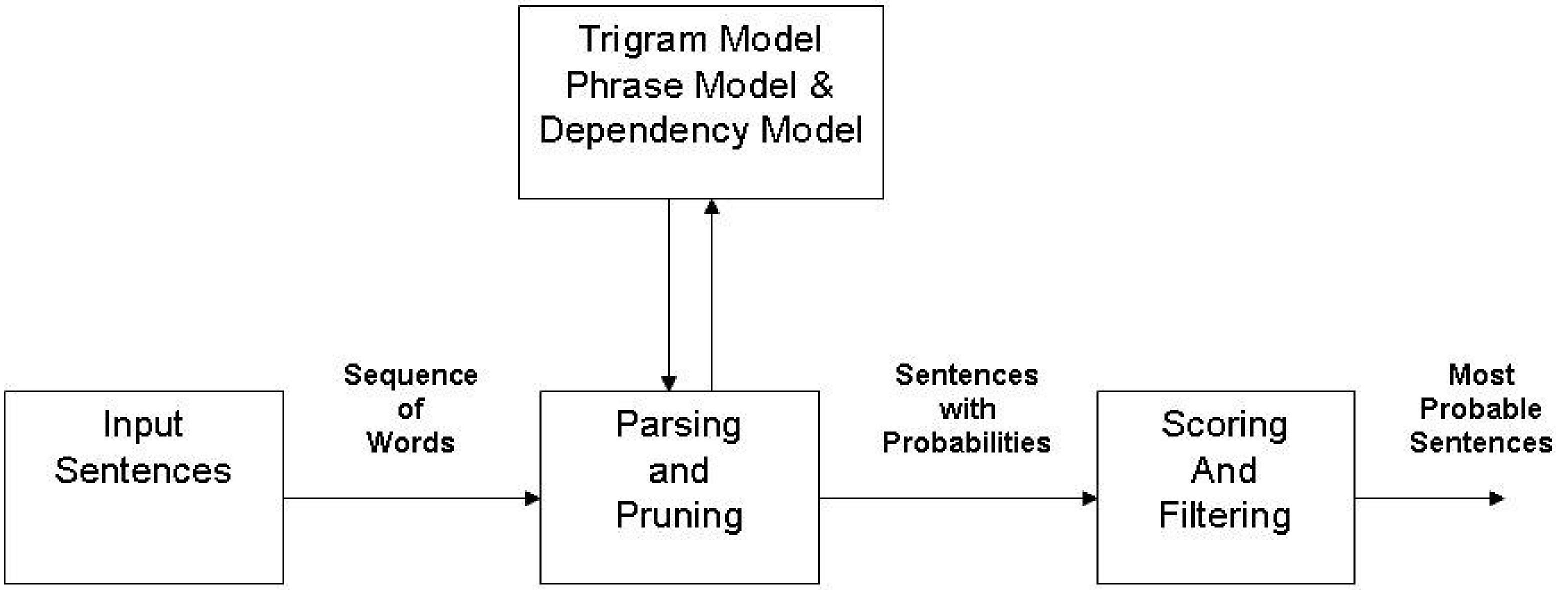


Figure: Lexicalized Statistical Parsing

Source: https://forum.huawei.com/enterprise/en/what-is-parsing-in-nlp/thread/571685-100429

Lexicalized statistical parsing (2 of 2)

- Step 1: Lexicalization.
  - Remove the beginning and ending markers in a sentence.
  - Removal of special characters and punctuations.
  - Create a tree bank.

- Step 2: Language model construction.
  - Tree bank ➔ Phrase structure or dependency structure.
  - Tree bank ➔ Generate the features, probabilities of the words.
  - Model calculation ➔ Relations between the words.
  - Dependency association.

- Step 3: Statistical Parsing Implementation:
  - The syntax, semantics, relationship of words ➔ Parse tree.
  - Long-term relation ➔ Higher level through the complex structures.

Top-down parsing

- The top down parsing method begins on the top with the start symbol "S“.

Example:

- Sentence: Maybe john walks

Grammar:

$s \Rightarrow sadv, s$

$s \Rightarrow np, vp$

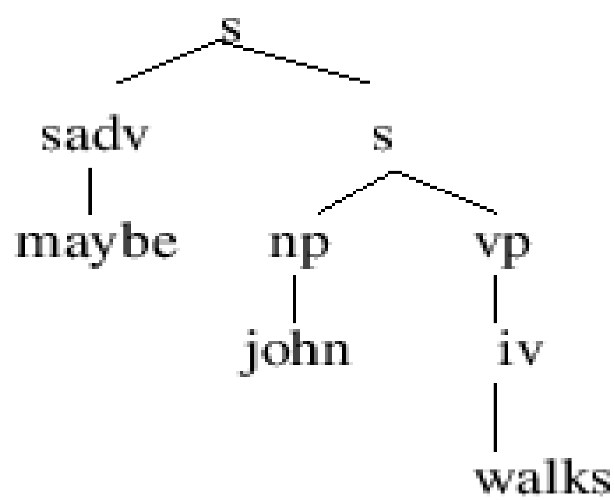
$np \Rightarrow$ john

$vp \Rightarrow iv$

$iv \Rightarrow$ walks

$sadv \Rightarrow$ maybe

Parse Tree



Bottom-up parsing

- The process starts from the non-terminal symbols and continuous upwards replacing the individual words in two sentence phrases until it reaches the root symbol.

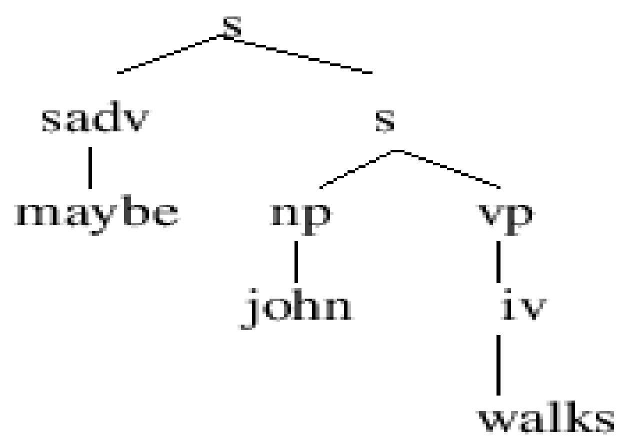
Example:

- Sentence: maybe john walks

Grammar:

$$s \Rightarrow sadv, s$$
$$s \Rightarrow np, vp$$
$$np \Rightarrow \text{john}$$
$$vp \Rightarrow iv$$
$$iv \Rightarrow \text{walks}$$
$$sadv \Rightarrow \text{maybe}$$

Parse Tree



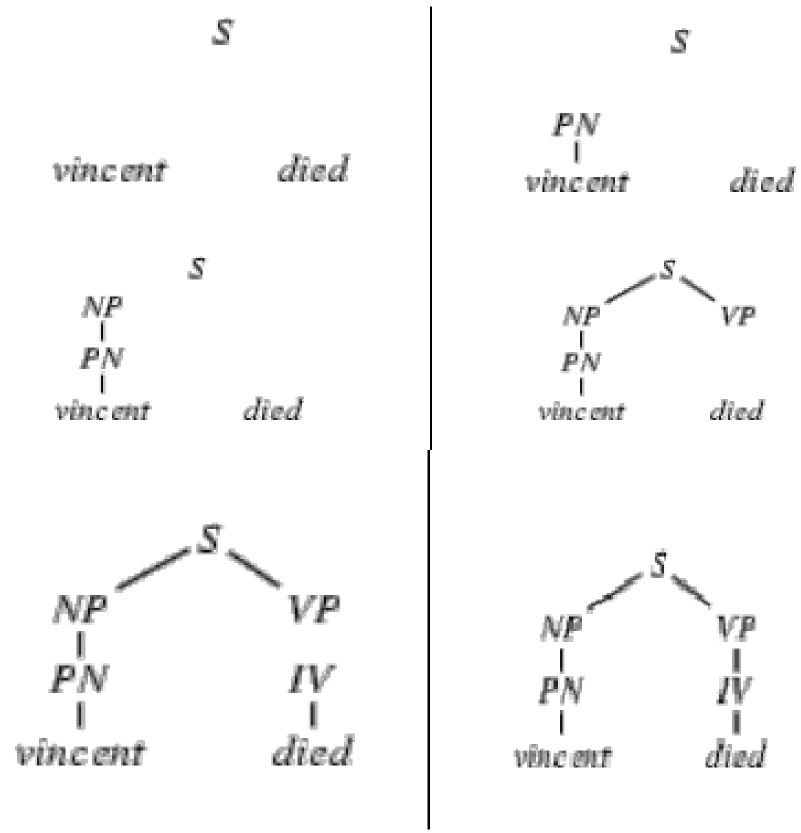
Left corner parsing method

- Example 1: Input Statement: vincent died.

Steps:

- Input: vincent Recognize an s. (Top-down prediction.)
- First word ➔ pn. (Bottom-up)
- pn at left corner: np → pn. (Bottom-up)
- np at left corner: s → np vp (Bottom-up)

- LHS = RHS.
- Input: died. Recognize a vp. (Top-down.)
- First word ➔ iv. (Bottom-up.)
- iv at left corner: vp → iv. (Bottom-up.)
- LHS = RHS



Statistical parsing: Probabilistic parser

- Probabilistic CFG (PCFG).

- A CFG in which its re-writing rules are associated with a probability. p ➔ Probability of a non-terminal A expanded to sequence β.

$$A \rightarrow \beta\ [p]$$

- Probability of an expansion RHS β given the LHS A.

$$\mathrm{P}\ (\mathrm{RHS}\ |\mathrm{LHS}\ )$$

S -> NP VP [.80]
S -> Aux NP VP [.15]
S -> VP [.05]
NP -> Pronoun [.35]
NP -> Proper-Noun [.30]
NP -> Det Nominal [.20]
NP -> Nominal [.15]

Figure: PCFG Representation

Source: http://disi.unitn.it/~bernardi/Courses/CL/Slides/9_stat_parsing.pdf

Multiword expressions

- Multi word expressions are Idiosyncratic.
- The word that syntactically or semantically similar.

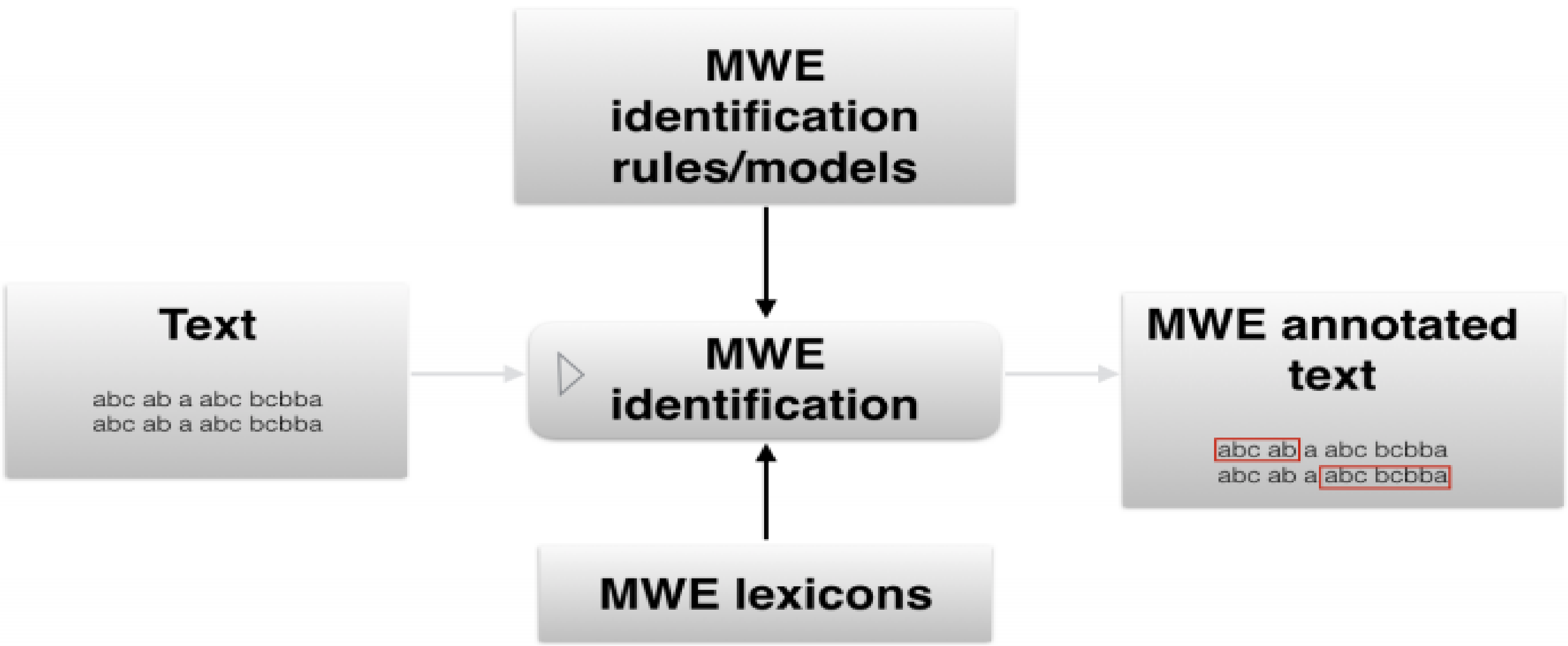


Figure: Multiword Expressions Process Outline

Source: https://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00302

Features of MWE

- The MWE Can be decomposed into multiple lexemes.

- These can be represented through syntactic, semantic, pragmatic or lexical units.

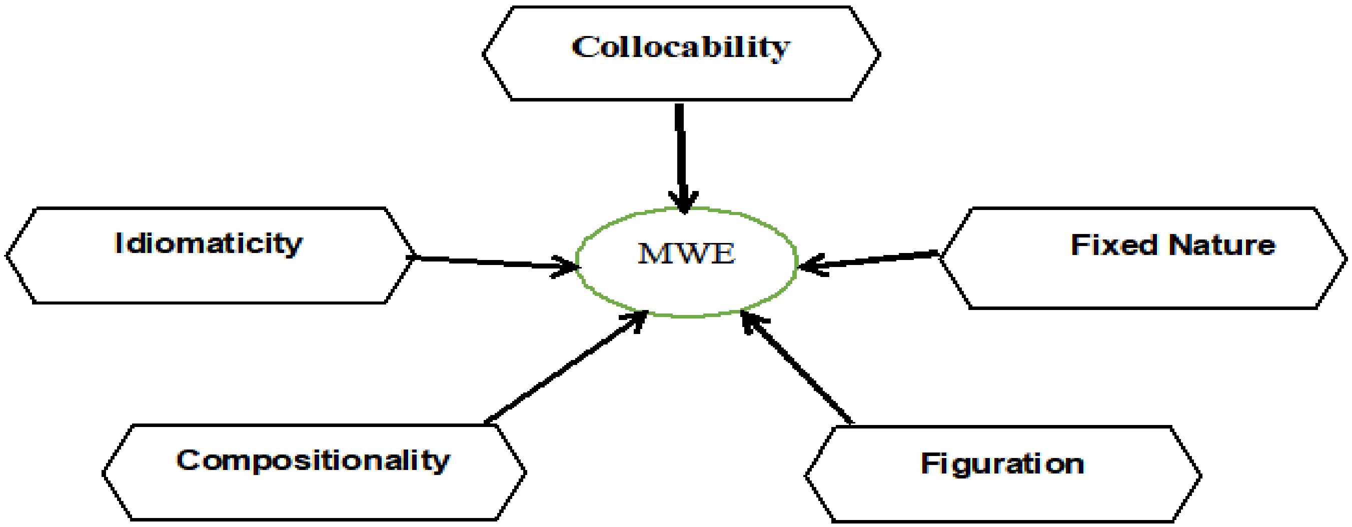


Figure: MWE Features

.

Types of multi word expressions

Atkins and Rundell (2008)	Bergenholtz and Gouws (2014)	Baldwin and Kim (2010)
I. COLLOCATIONS		
collocations risk one's life	collocations severe criticism	collocations immaculate performance
II. FIXED PHRASES & IDIOMS		
phrasal idioms to have a heart of gold	idioms to have eyes in the back of one's head	verb-noun idiomatic – combinations kick the bucket
fixed phrases ham and eggs	non-pictorial idiomatic MWE round the clock	
similes drunk as a lord	twin formula day and night	
	comparative MWE as right as rain	
	MWEs from foreign languages ad hoc	
	(non)idiomatic MWEs with a unique component to and fro	
	MWEs with an old inflection	
III. COMPOUNDS		
figurative compounds lame duck	semi-terms magic eye	nominal compounds golf club, connecting flight
semi-figurative compounds high school		
functional compounds police dog		
IV. PROVERBS		
proverbs too many cooks …	proverbs half a loaf is better than no bread	sentence-like units good-morning
quotations to be or not to be	winged words One small step for man …	
greetings good morning	routine formulas how do you do	
phatic phrases have a nice day	expletive constructions give him an inch and …	
catch phrases horses for courses		
V. PHRASAL VERBS		
phrasal verbs get up, see trough	non/idiomatic particle verb to run at/to bask in	verb-particle constructions take off
	non/idiomatic reflexive verb to enjoy yourself/to prostitute yourself	prepositional verbs refer to
VI. LIGHT-VERB CONSTRUCTIONS		
support verb constructions to take a decision	noun phrase with semantically void verb set in motion	light-verb constructions to take a walk
VII. PREPOSITIONAL PHRASES		
compound prepositions in spite of	MWEs with syntactic function with regard to	prepositional phrases in bed, in jail
		complex prepositions on top of

Multi word verbs

- The verbs that contain more than one word are called as multi word verbs.

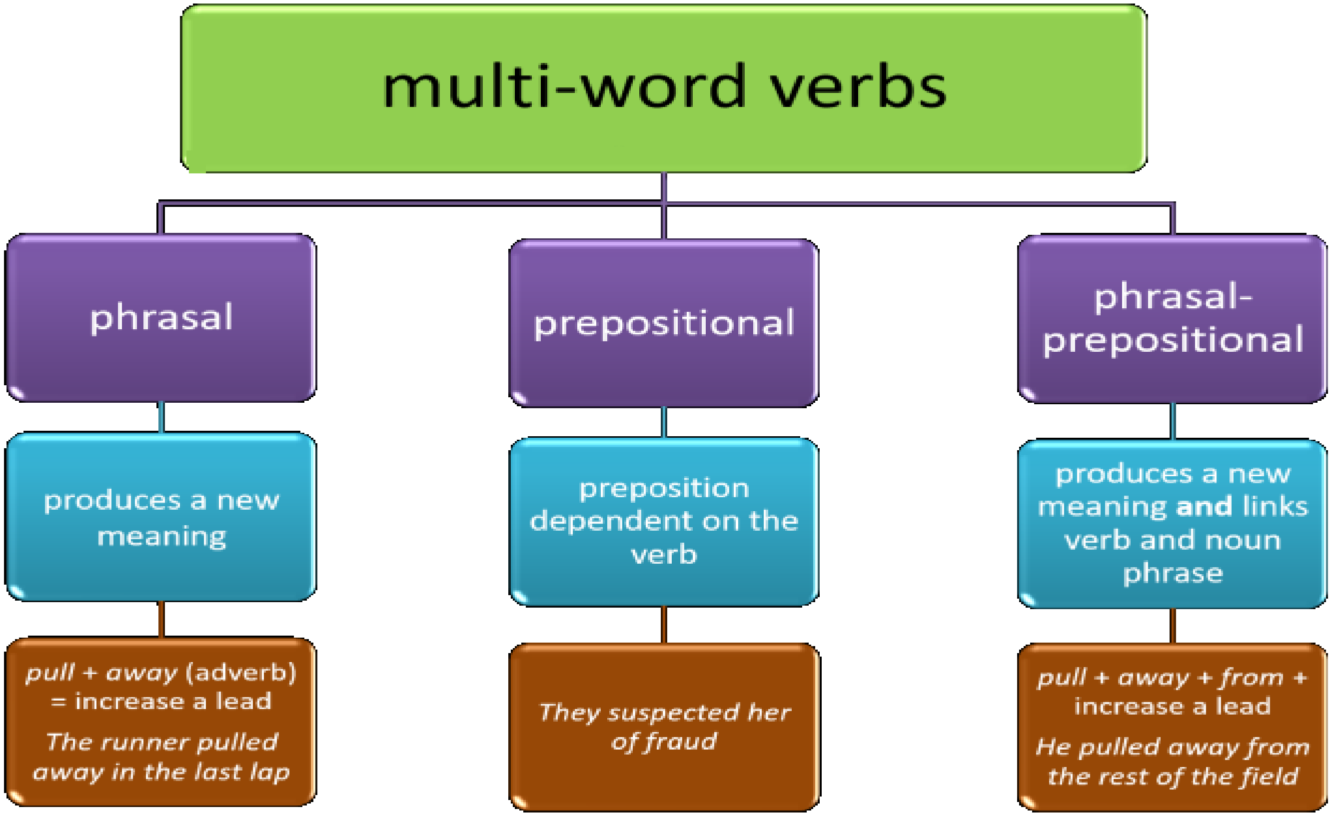


Figure: Multi Word Verbs

Source: https://www.eltconcourse.com/training/inservice/verbs/mwvs.html

Word similarity and text similarity

- Helps in determining the closeness of two or more words/text.

- bag of words, TF-IDF, word2vec etc. are used to encode the input text data.

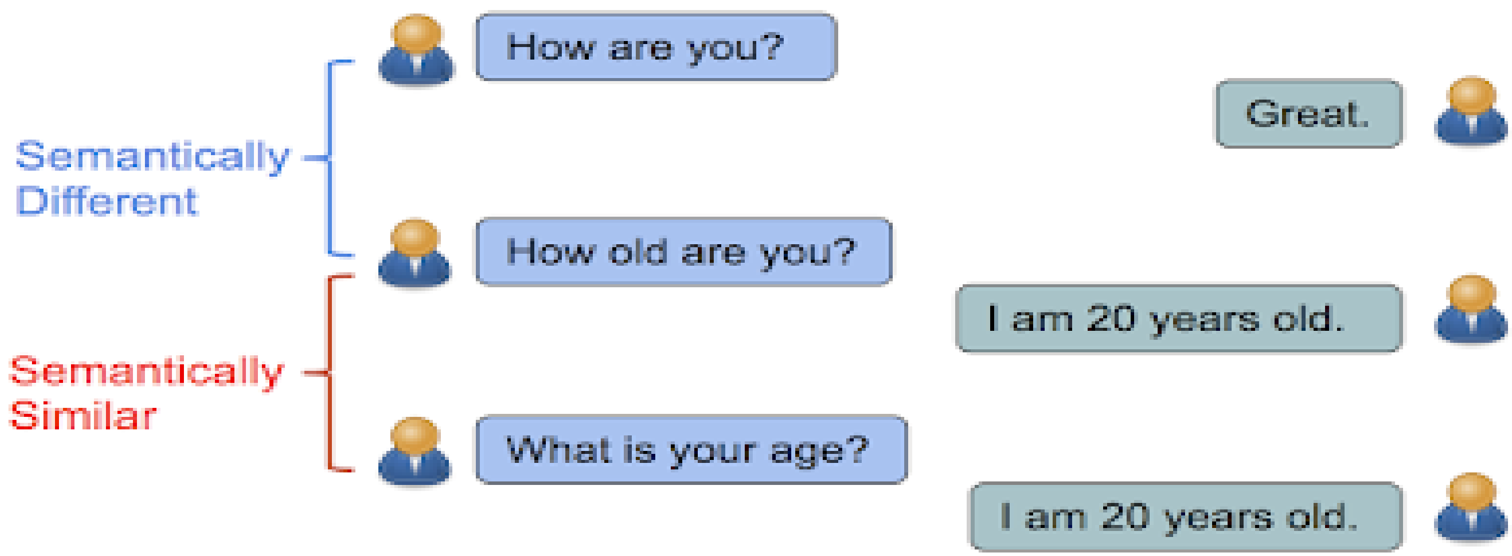


Figure: Semantic similarity and dissimilarity

Source: https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html

Normalized web distance

- Relevant content ➔ Checked for similarity.
- Normalized web distance ➔ Similarity between words.
- Query engines ➔ Normalized web distance ➔ Aggregate results.
- Higher similarity ➔ Stacked on the top of the results.
- NWD method identifies the semantic relations between arbitrary objects.
- Parameter free and feature free data mining method.

- Methods for word similarity:
  - Association measures.
  - Attributes.
  - Relational word similarity.
  - Latent semantic analysis.

- Applications:
  - Hierarchical clustering.
  - Classification.
  - Matching the meaning.
  - Systematic comparison.

Text similarity methods

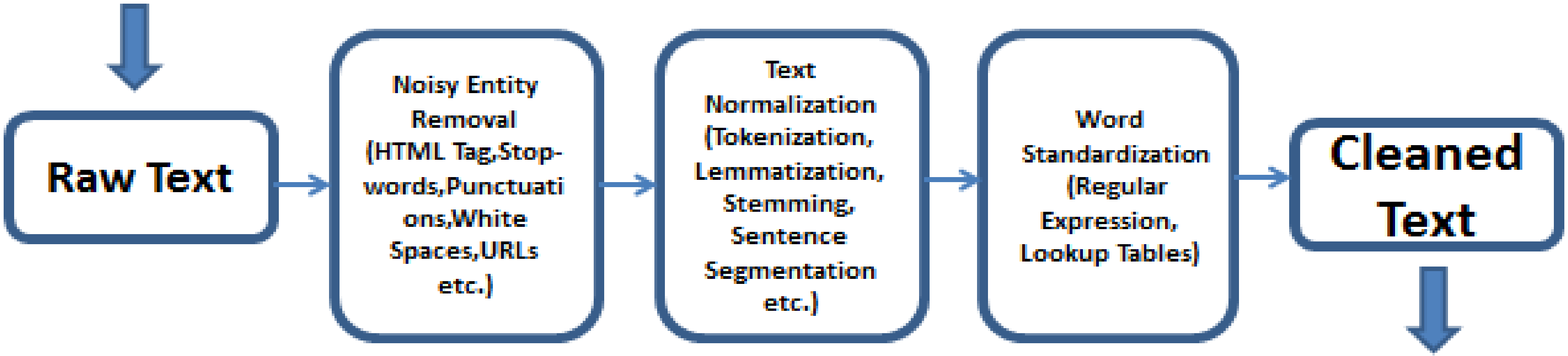


Figure: Pre-Processing of Text

Source: http://www.vanaudelanalytix.com/python-blog/pre-processing-text-for-nlp

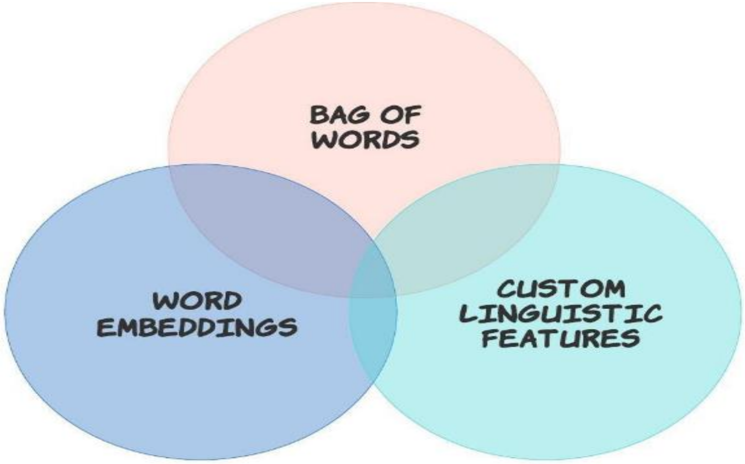


Figure: Feature extraction

Source: https://amp.flipboard.com/@tdatascience/artificial-intelligence-8qhakstrz/the-triune-pipeline-for-three-major-transformers-in-nlp/a-WXncOskwRTGZbgY5j66HcA%3Aa%3A2892075988-6fae262963%2Ftowardsdatascience.com

Jaccard similarity

- Simple representation of two text sentences that have common elements intersection over Union method.

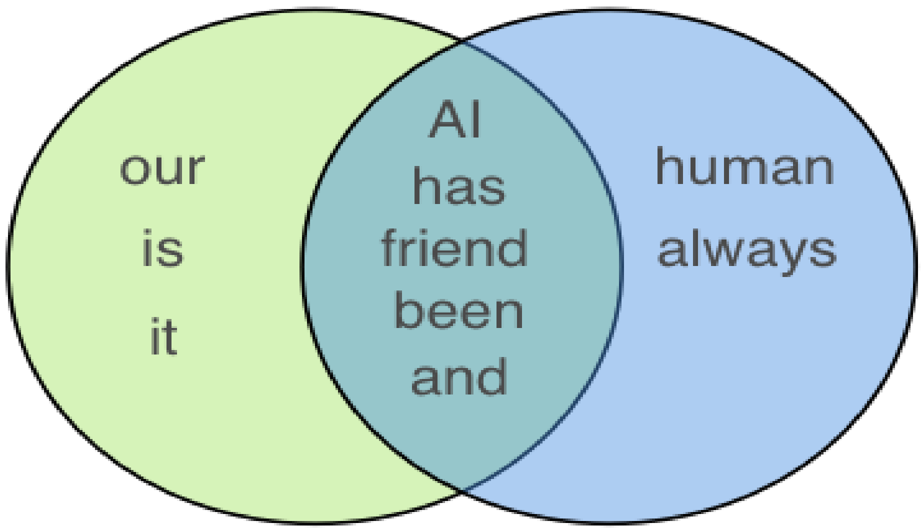


Figure: Jaccard distance

Source: https://medium.com/@adriensieg/text-similarities-da019229c894

K-means

- Usage of K means algorithm for conversion of the words into appropriate vector.

	Document	Category	ClusterLabel
0	The sky is blue and beautiful.	weather	2
1	Love this blue and beautiful sky!	weather	2
2	The quick brown fox jumps over the lazy dog.	animals	1
3	A king's breakfast has sausages, ham, bacon, eggs, toast and beans	food	3
4	I love green eggs, ham, sausages and bacon!	food	3
5	The brown fox is quick and the blue dog is lazy!	animals	1
6	The sky is very blue and the sky is very beautiful today	weather	2
7	The dog is lazy but the brown fox is quick!	animals	1
8	President greets the press in Chicago	politics	4
9	Obama speaks in Illinois	politics	4

Figure: K – Means

Source: https://medium.com/@adriensieg/text-similarities-da019229c894

Cosine similarity

- Uses the cos angle between the vectors.

- Measure of similarity between two non-zero vectors based on the inner product of cosine angles.

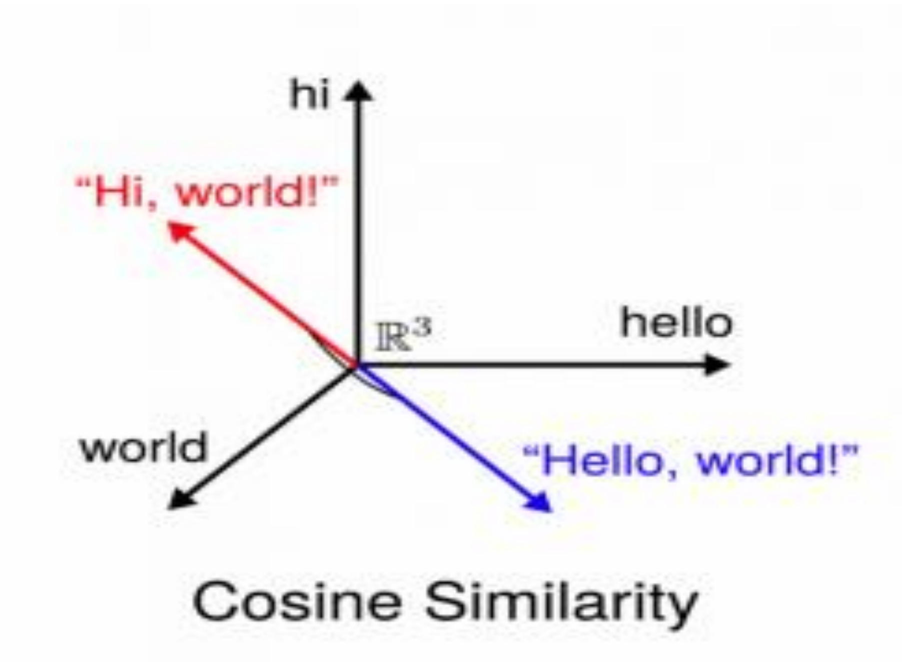


Figure: Cosine Similarity

Source: https://medium.com/@adriensieg/text-similarities-da019229c894

Word Mover's distance

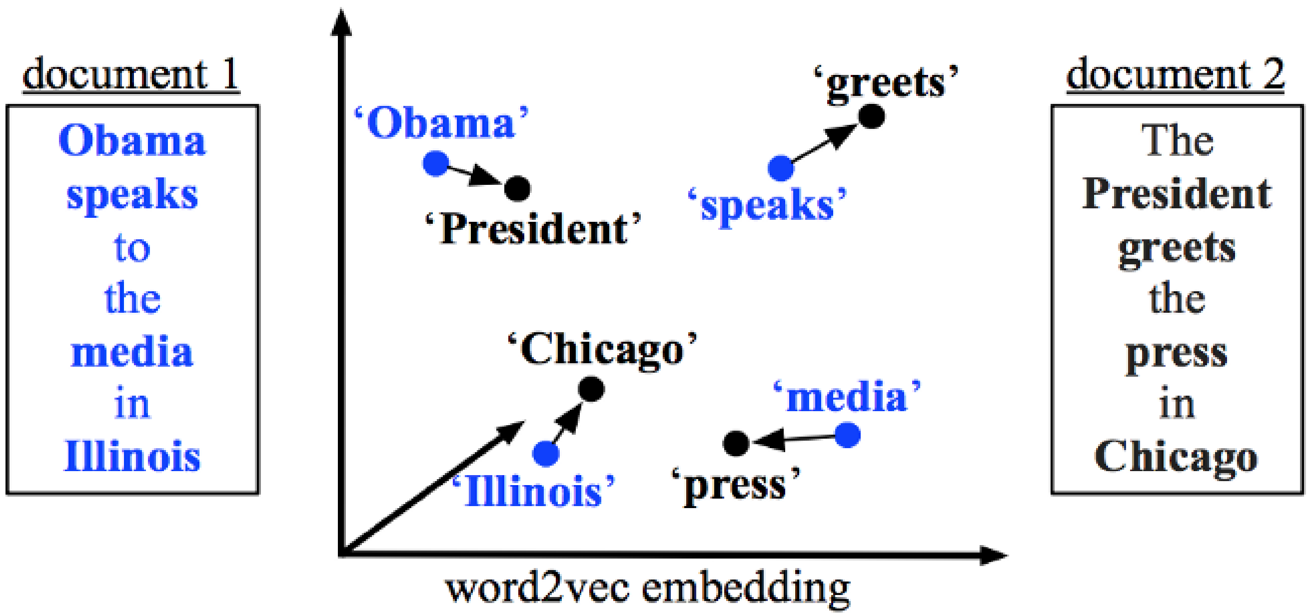


Figure: Word Mover's Distance

Source: https://medium.com/@adriensieg/text-similarities-da019229c894

Variational auto encoders

- Used to identify text based upon the same text as input.
- The auto-encoder uses neural network to an approximate value relative to the input.

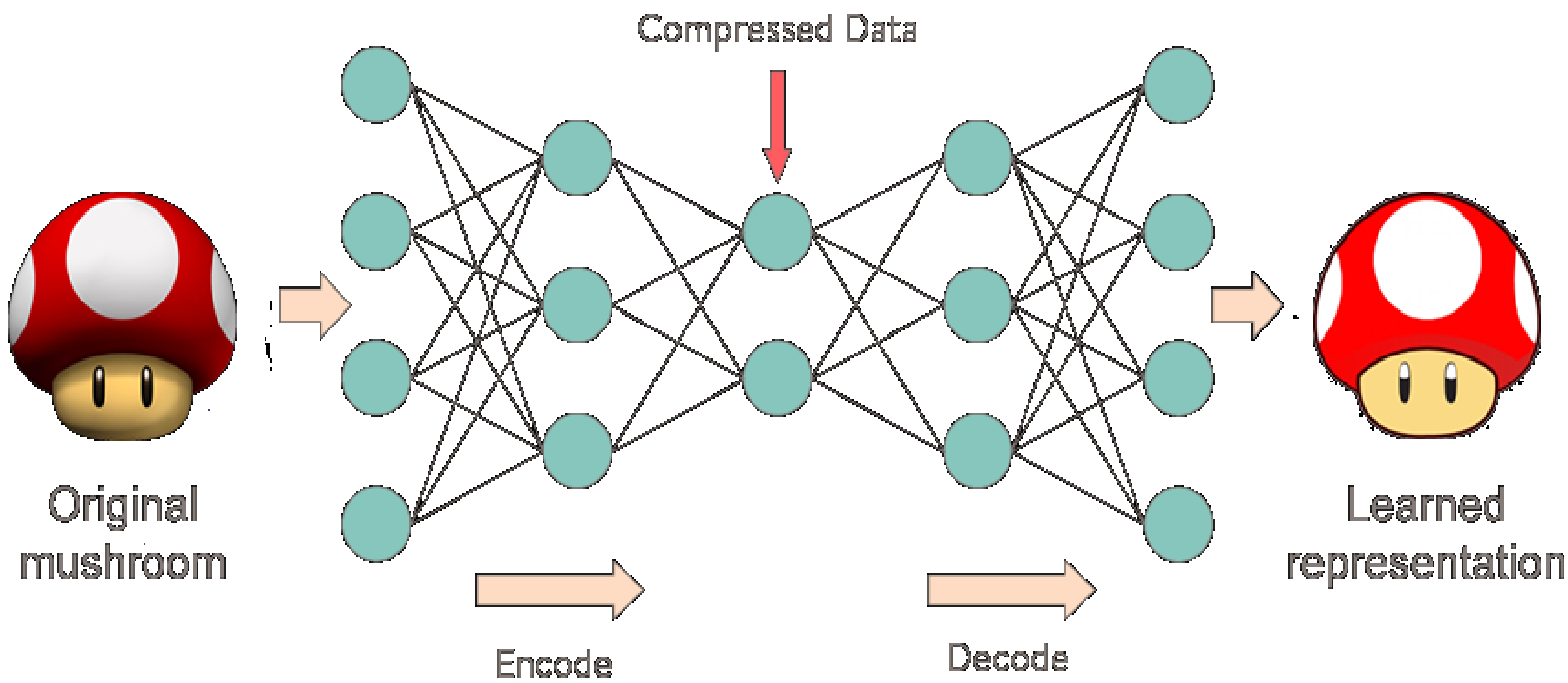


Figure: Variational Auto Encoders

Source: https://medium.com/@adriensieg/text-similarities-da019229c895

Pre-trained sentence encoders

- Used to encode the basic text into higher dimension vectors.

- Pre-trained encoders are trained on both supervised learning and unsupervised learning to identify both the syntactic and semantic information.

Figure: Pre-Trained Sentence Encoders

Source: https://medium.com/@adriensieg/text-similarities-da019229c896

Bidirectional Encoder Representations from Transformers (BERT) with cosine distance

- The BERT model uses word vectors that can adapt depending upon the surroundings.

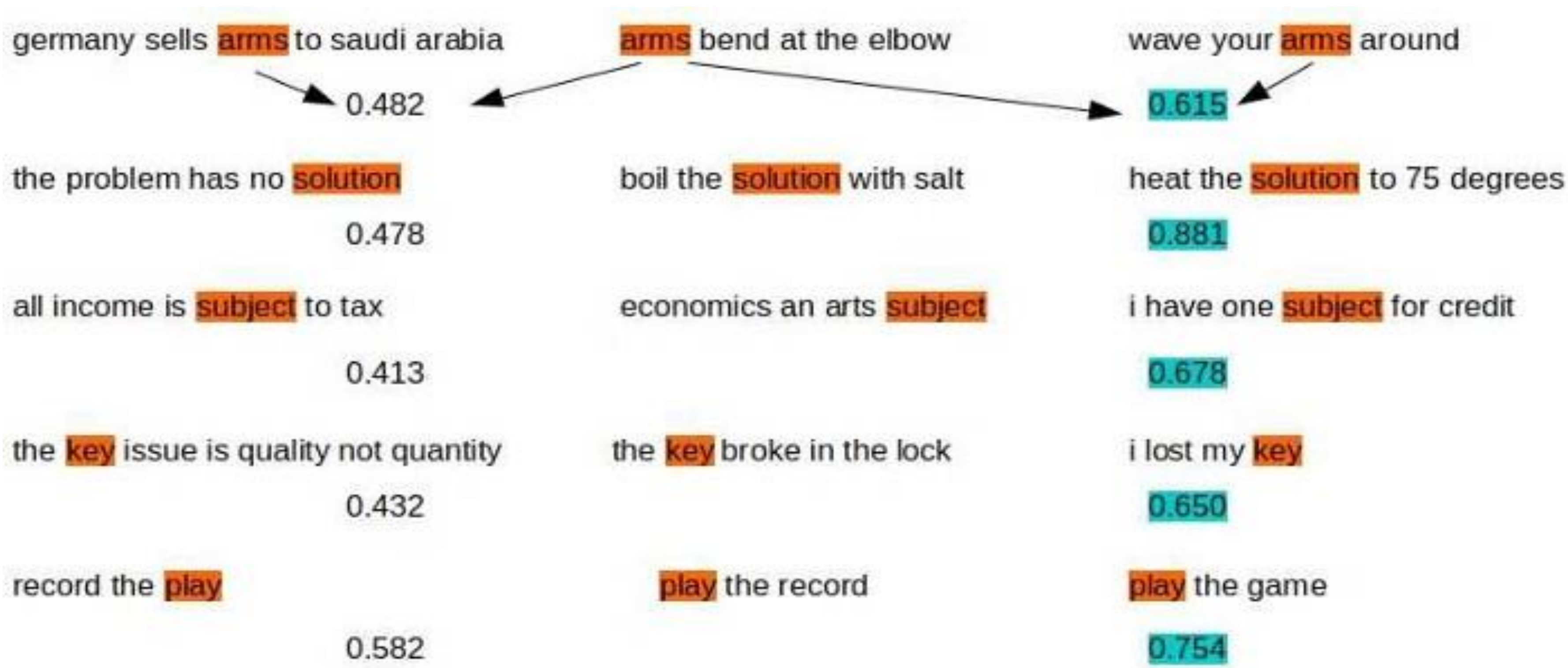


Figure: BERT Similarity

Source: https://medium.com/@adriensieg/text-similarities-da019229c897

Self evaluation: Exercise 9

- To continue with the training, after learning the concepts of Parsing, Tokenization, Stop Word Removal in Natural Language Text Processing, it is time to write code to work with Tokenization, Tagging, Parsing and use it. It is instructed to utilize the concepts of reading data from files Tokenization, POS Tagging, Parsing and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 9: Read any text, perform tokenization and POS tagging as a preprocessing activity. Create parse trees from the preprocessed text and display them. Use Treebank chunks from NLTK corpus and create Parse tree on the same. Draw the Parse trees also.

Self evaluation: Exercise 10

- To continue with the training, after learning the concepts of Parsing, Tokenization, Stop Word Removal in Natural Language Text Processing, it is time to write code to work with Tokenization, Tagging, Parsing and use it. It is instructed to utilize the concepts of reading data from files Tokenization, POS Tagging, Parsing and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 10: Python code to Read two set of documents, perform Tokenization, map dictionaries, create corpus and calculate the similarity between the documents.

Word sense disambiguation

- Word sense disambiguation related to identify the sense of any word use tree in a sentence.
- Identify and determine the meanings of the words in any context.

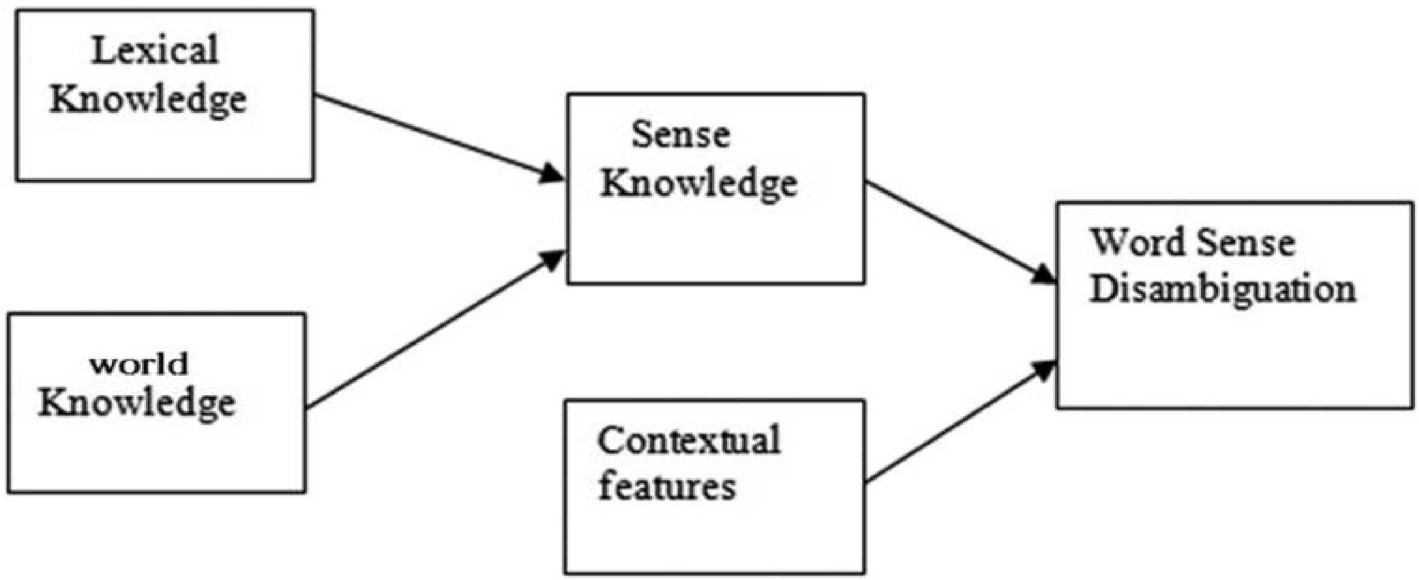


Figure: Word Sense Disambiguation Process

Source: https://content.iospress.com/articles/international-journal-of-knowledge-based-and-intelligent-engineering-systems/kes190399

Complications in WSD

- Dictionary differences: Word identified with appropriate senses.

- POS tagging: Positioning of the words in the sentence to understand the tag and the meaning.

- Inter-judge variance: Human sensing of words and their meaning ➔ Hard to decipher.
  - Sense cannot be same for all.

- Pragmatics: Identifying the meaning of the context based upon ontology.

- Example:
  - Sentence 1: Alex and John are fathers - Independent relationship.
  - Sentence 2: Alex and John are brothers - Dependent relationship.

Methods in WSD

- Deep approach.
- Shallow approach.

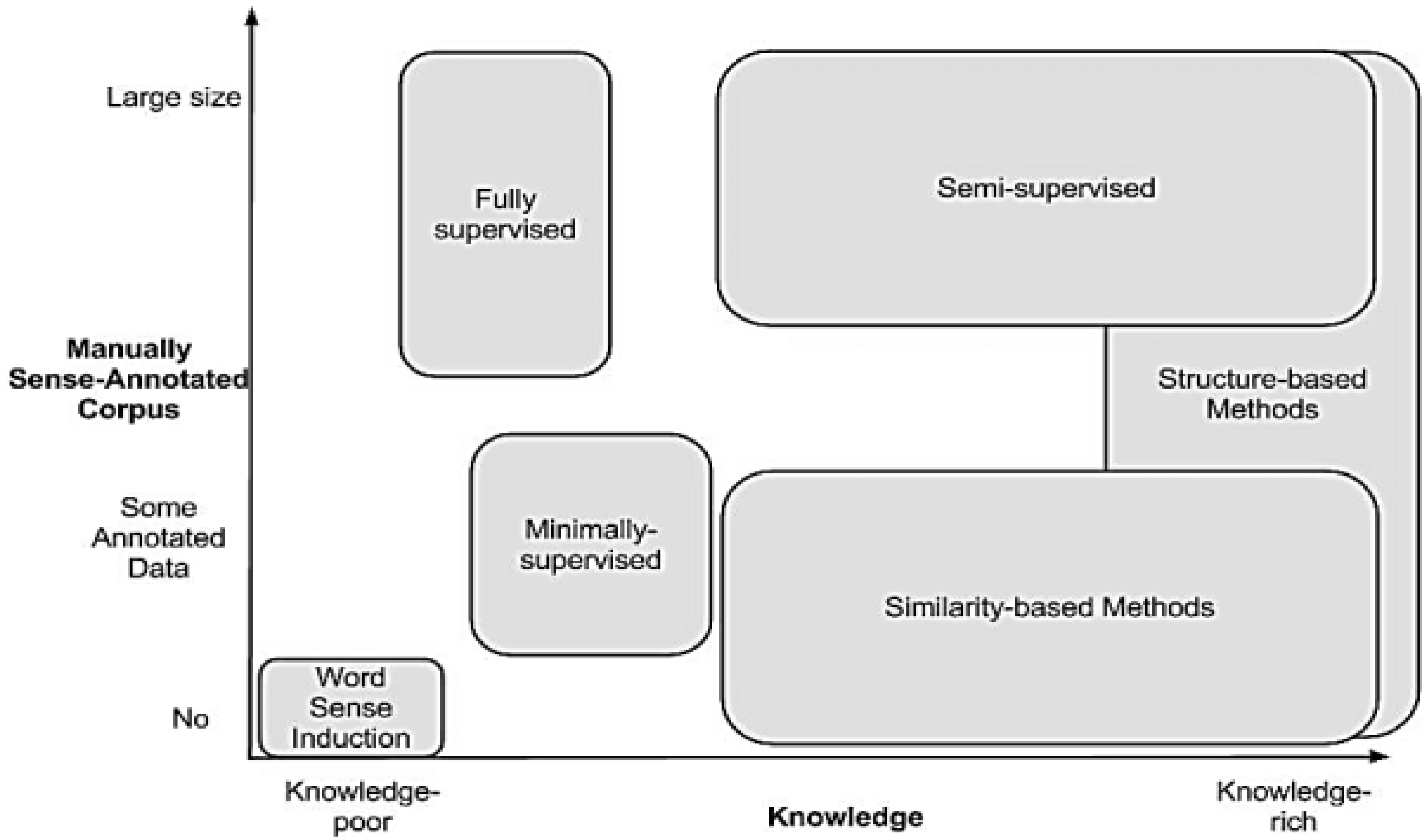


Figure: Methods in WSD

Source: https://www.researchgate.net/figure/Word-Sense-Disambiguation-systems-Data-versus-Knowledge-Schwab-2013-personal-notes_fig2_257409694

Evaluation of WSD

- Very hard to evaluate ➔ Every word can have different sense based upon the context.
- Requires large amount of hand annotated Corpus.

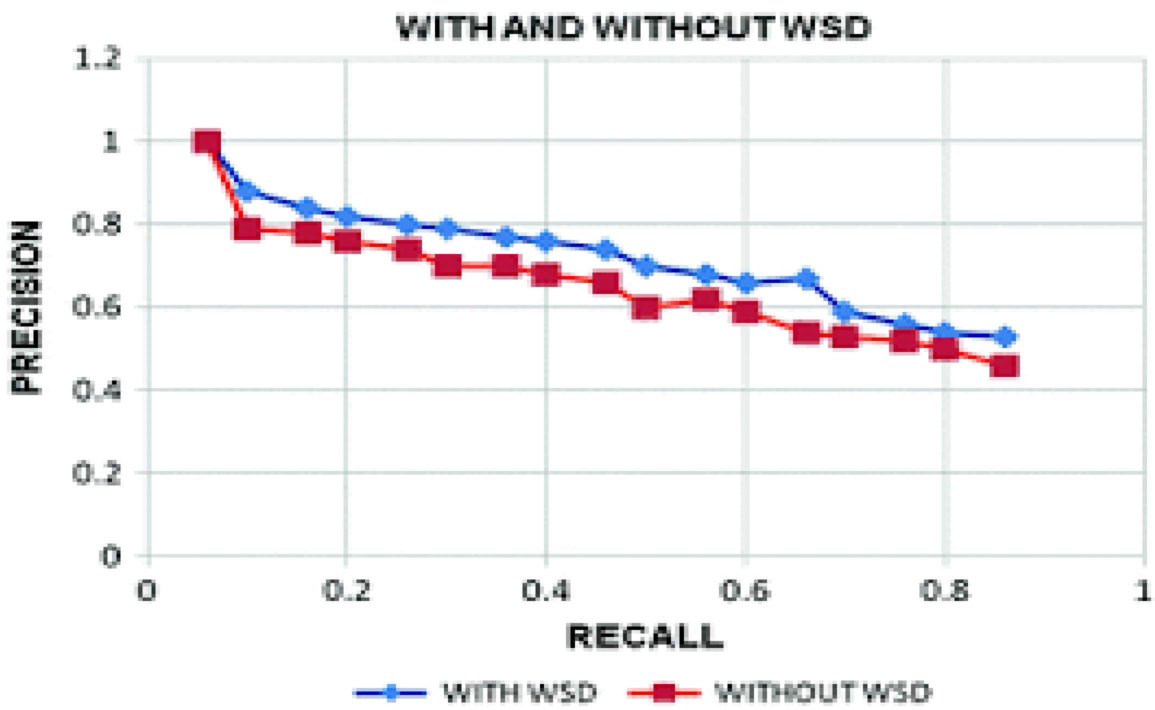


Figure: Evaluation of WSD

Source: https://link.springer.com/chapter/10.1007/978-981-10-2471-9_64

Self evaluation: Exercise 11

- To continue with the training, after learning the concepts of Parsing, Tokenization, Word Sense Disambiguation, Stop Word Removal in Natural Language Text Processing, it is time to write code to work with Tokenization, WSD and use it to compare similarities. It is instructed to utilize the concepts of reading data from files Tokenization, Word Similarity, WSD and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 11: Read words and sentences and perform Word Sense Disambiguation using WordNet and LESK.

Self evaluation: Exercise 12

- To continue with the training, after learning the concepts of Parsing, Tokenization, Word Sense Disambiguation, Stop Word Removal in Natural Language Text Processing, it is time to write code to work with Tokenization, WSD and use it to compare similarities. It is instructed to utilize the concepts of reading data from files Tokenization, Word Similarity, WSD and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 12: Read text from files, perform pre-processing activities like Word Sense Disambiguation, tokenization, stop word removal. Create a question answer context where the program can read queries from the user and respond as per the words and sentences in the question.

History of speech recognition technology

IBM ICE (Innovation Centre for Education)

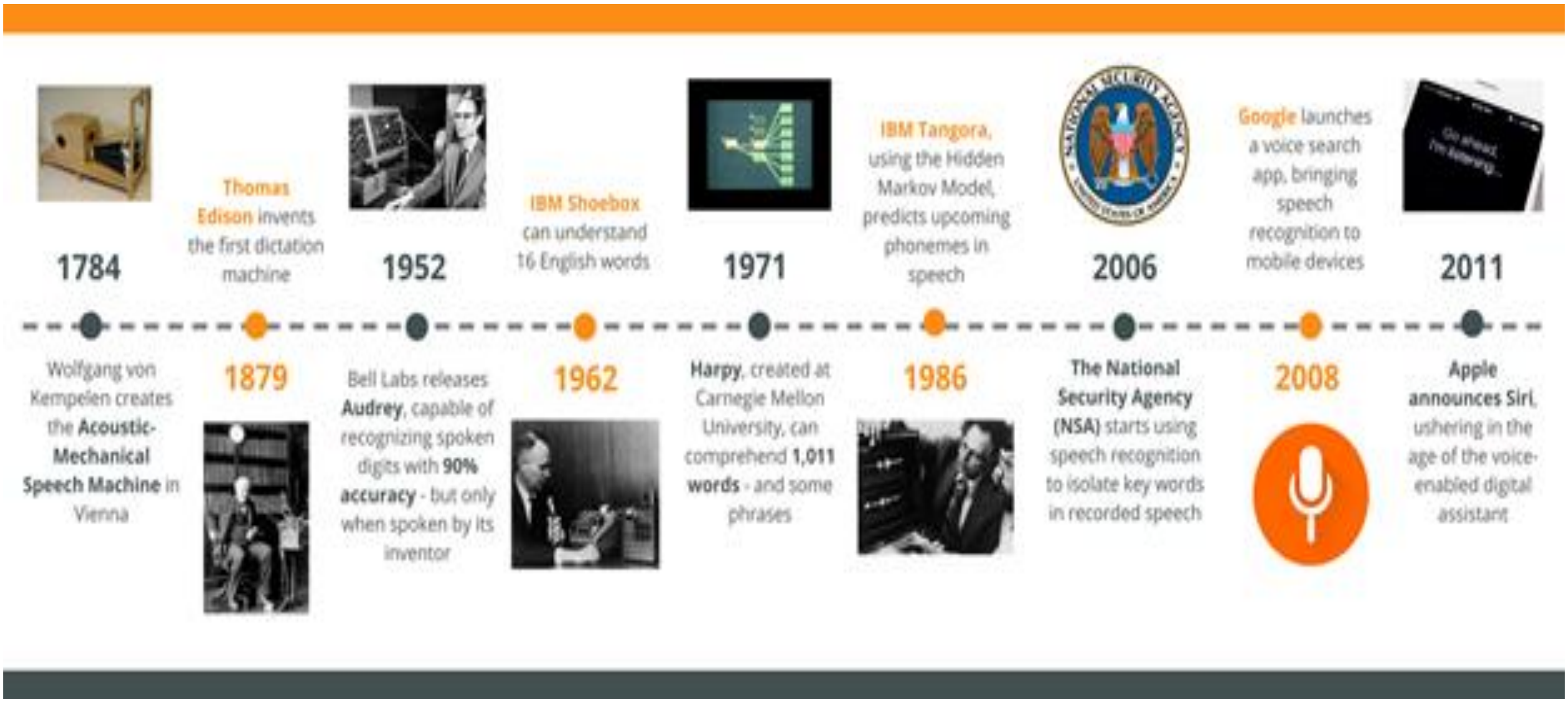


Figure: History of Speech Recognition Technology

Source: https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf

Working principle in voice recognition

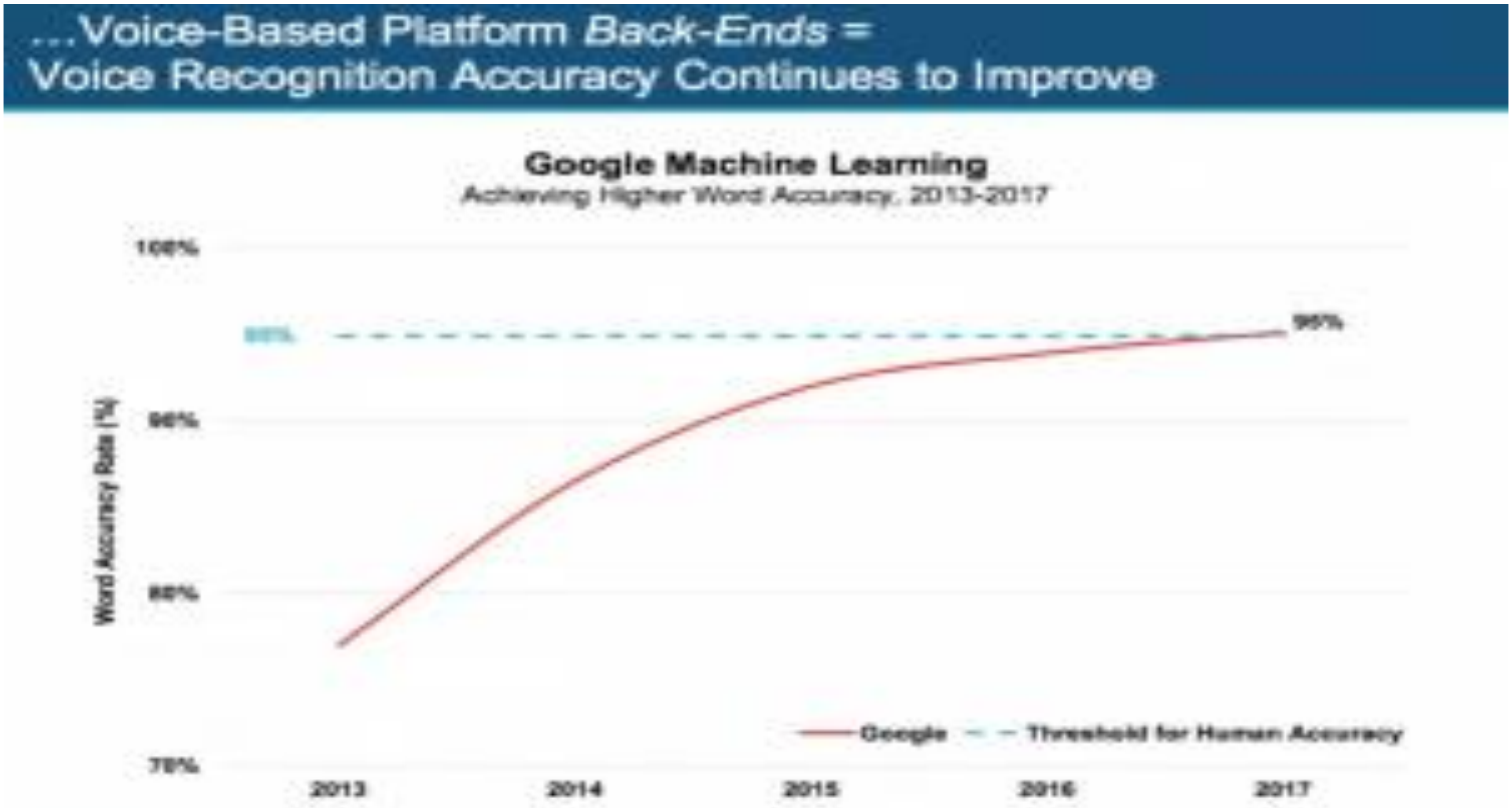


Figure: Voice Recognition Accuracy

Source: https://www.globalme.net/blog/the-present-future-of-speech-recognition/

Major leaders in speech recognition and voice assistant

- Apple's Siri.
- Hardware: Apple HomePod (Due to launch in 2018 at $349), iPhone, MacBooks, AirPods.
- Digital assistant: Siri
- Usage statistics:
  - 42.5% of smartphones have Apple's Siri digital assistant installed (Highervisibility).
  - 41.4 million monthly active users in the U.S. as of July 2017, down 15% on the previous year (Verto Analytics).
  - 19% of iPhone users engage with Siri at least daily (HubSpot).

Figure: Apple's Siri

Source: https://osxdaily.com/2016/05/03/improve-hey-siri-voice-training-ios/

Amazon Alexa

Amazon Alexa

- Hardware: Echo, Echo Dot, Echo Show, Fire TV Stick, Kindle..
- Digital Assistant: Alexa
- Usage Statistics:
  - "Tens of millions of Alexa-enabled devices" sold worldwide over the 2017 holiday season (Amazon).
  - 75% of all smart speakers sold to date are Amazon devices (Tech Republic).
  - There are now over 25,000 skills available for Alexa (Amazon).

Figure: Amazon Alexa

Source: https://www.imore.com/how-improve-amazon-alexa-voice-recognition

Microsoft Cortana

Microsoft Cortana

- Hardware: Harman/kardon Invoke speaker, Windows smartphones, Microsoft laptops
- Digital Assistant: Cortana
- Usage Statistics:
  - 5.1% of smartphones have the Cortana assistant installed
  - Cortana now has 133 million monthly users (Tech Radar)
  - 25% of Bing searches are by voice (Microsoft).

Figure: Cortana

Source: https://www.pcworld.com/article/2984791/how-to-enable-windows-10s-hey-cortana-voice-commands.html

Google Assistant

Google Assistant

- Hardware: Google Home, Google Home Mini, Google Home Max, Pixelbook, Pixel smartphones, Pixel Buds, Chromecast, Nest smart home products.
- Digital Assistant: Google Assistant.
- Usage Statistics:
  - Google Home has a 24% share of the US smart speaker market (eMarketer).
  - There are now over 1,000 Actions for Google Home (Google).
  - Google Assistant is available on over 225 home control brands and more than 1,500 devices (Google).

Figure: Google assistant

Source: https://www.indiatvnews.com/technology/news-google-assistant-devices-voice-match-default-speaker-know-what-is-it-625568

Machine translation

- Translate a text in one natural language to another natural language.

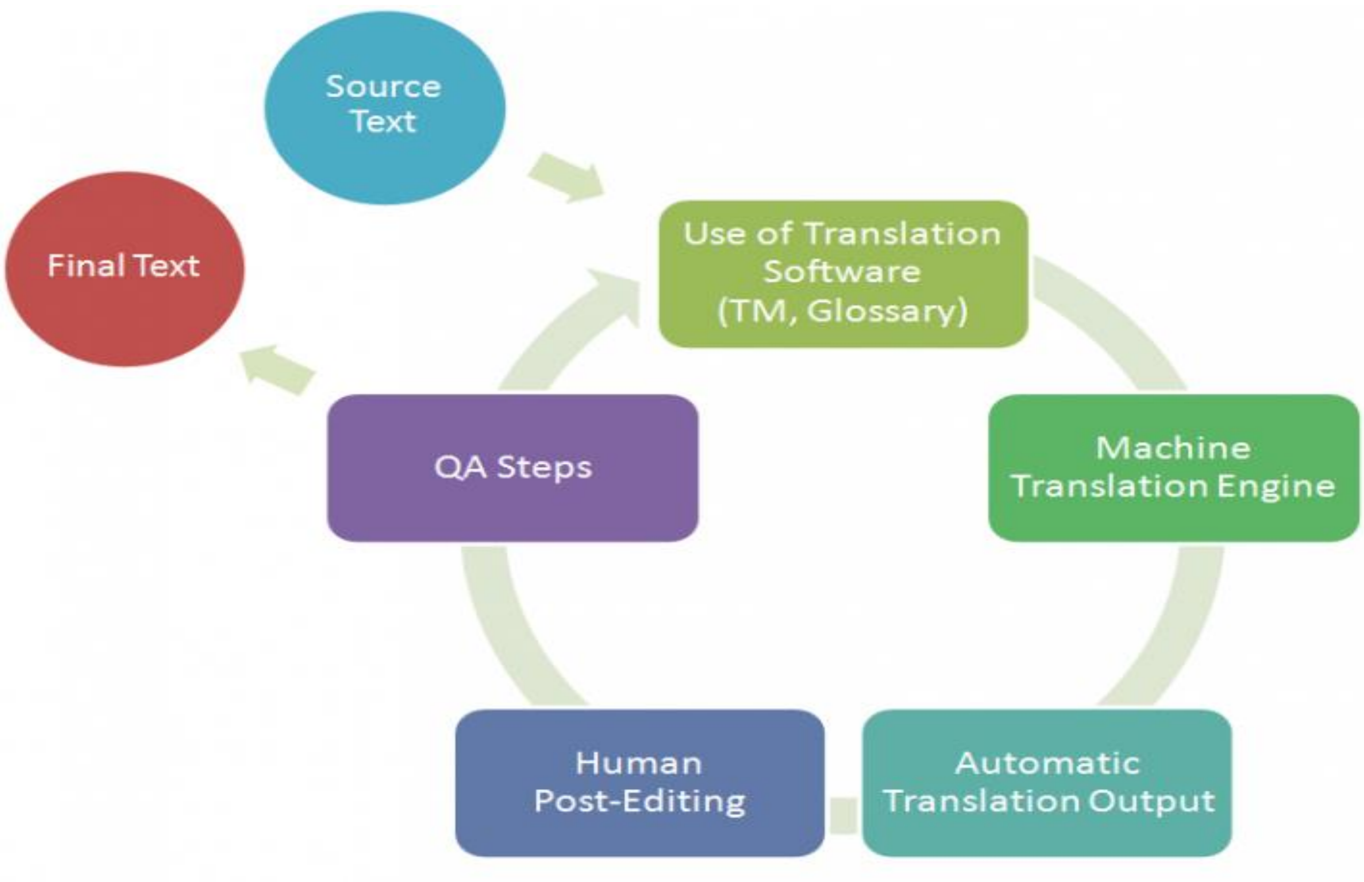


Figure: Machine Translation

Source: https://langsolinc.com/machine-translation-and-confidentiality/

Rule-based machine translation

- Parse through the text ➔ Create a representation of parse tree.
- Parse tree ➔ Text for the target language.
- Map linguistic universals (i.e., grammar) between languages.

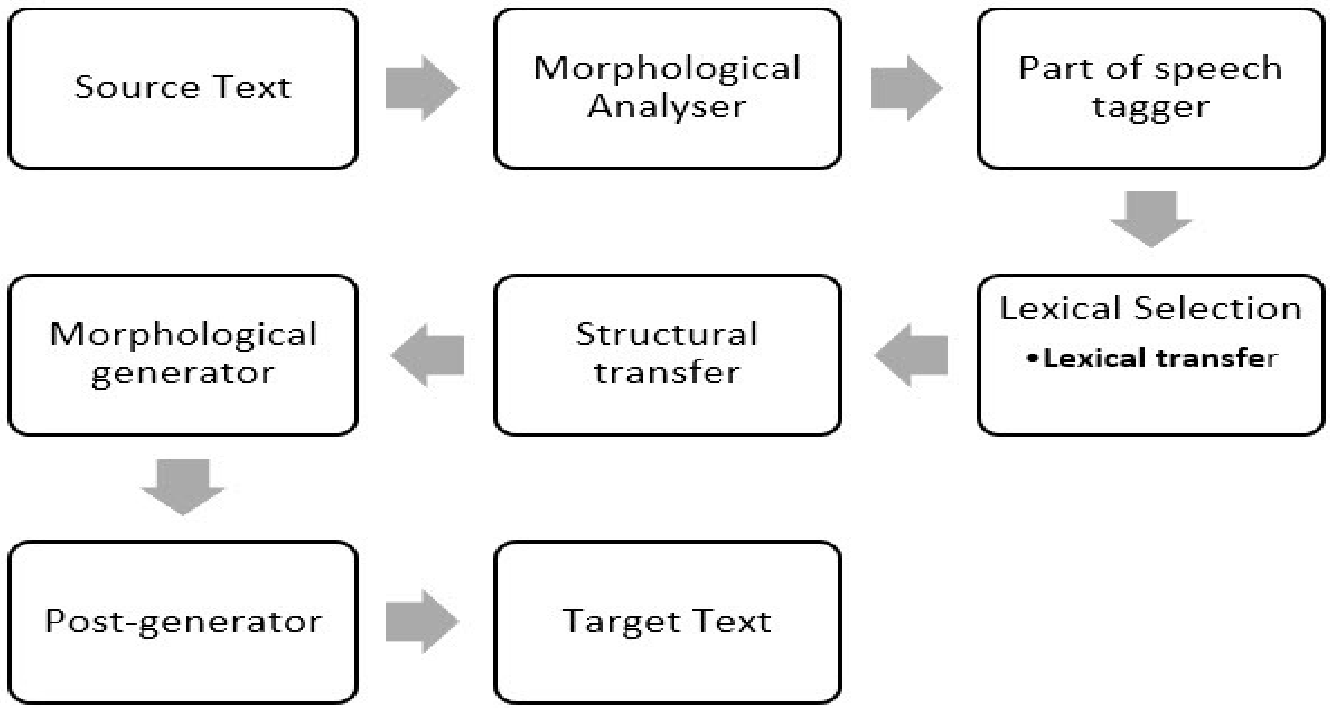


Figure: Rule-Based Machine Translation

Source: https://www.researchgate.net/figure/Rule-based-Machine-Translation_fig1_320730405

Statistical machine translation

- Language has an inherent logic that could be treated in the same way as any logical mathematical challenge.

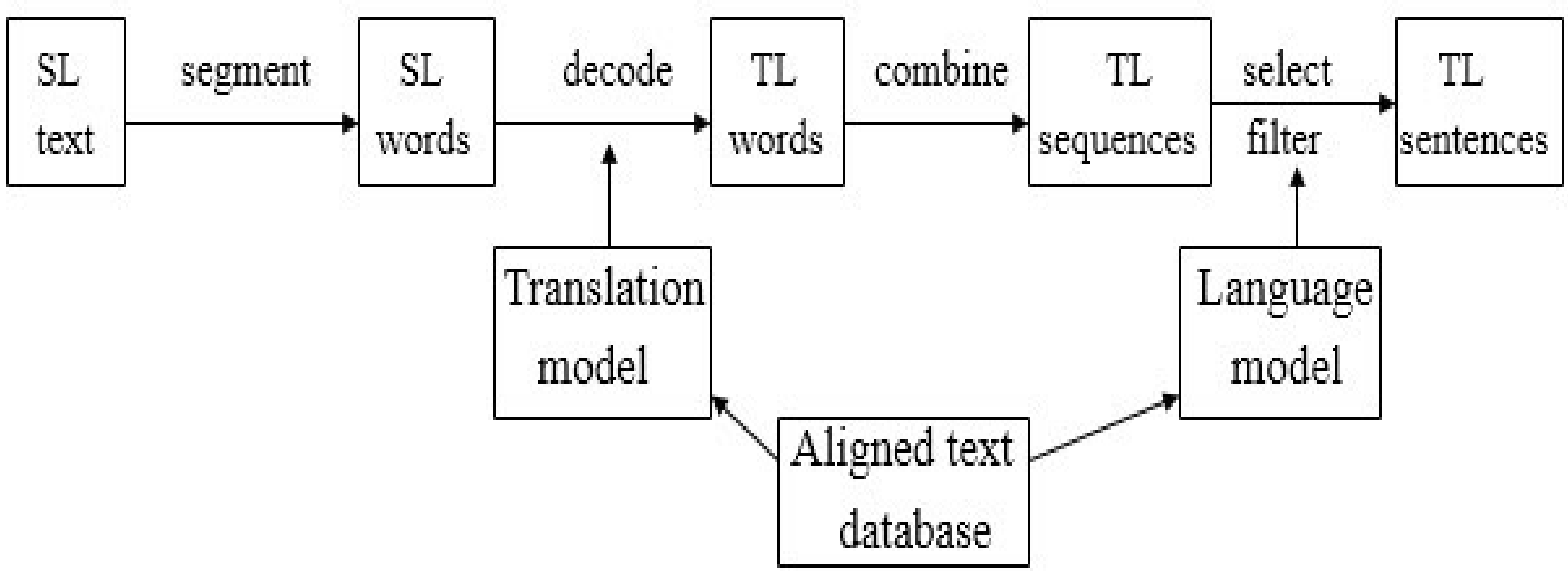


Figure: Statistical Machine Translation

Source: https://www.researchgate.net/figure/Statistical-Machine-Translation_fig2_320730405

Rule-based MT vs. statistical MT

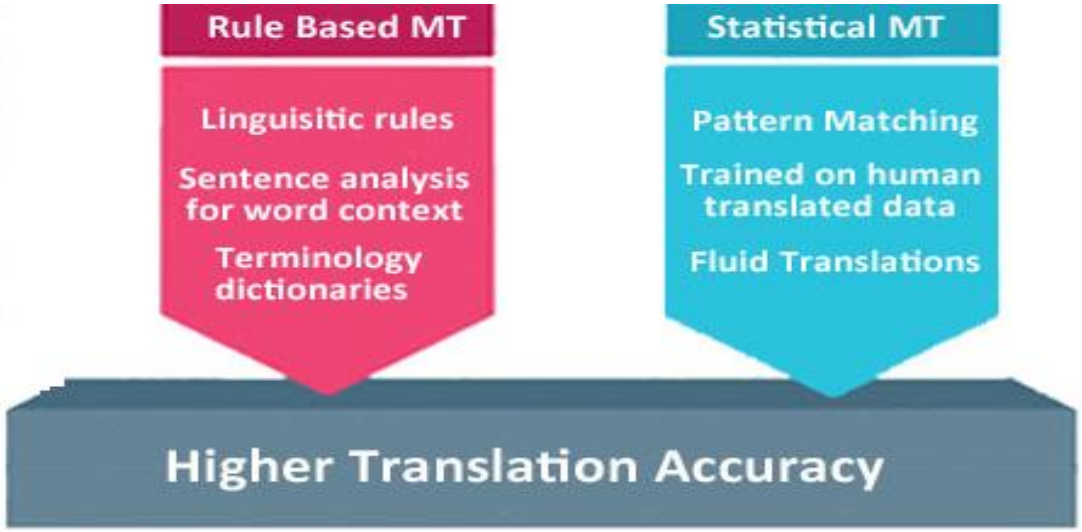


Figure: Rule-Based MT vs. Statistical MT

Source: https://www.translationsoftware4u.com/enterprise-global.php

Rule-Based MT	Statistical MT
Consistent and predictable quality	Unpredictable translation quality
Out-of-domain translation quality	Poor out-of-domain quality
Knows grammatical rules	Does not know grammar
High performance and robustness	High CPU and disk space requirements
Consistency between versions	Inconsistency between versions
Lack of fluency	Good fluency
Hard to handle exceptions to rules	Good for catching exceptions to rules
High development and customization costs	Rapid and cost-effective development costs provided the required corpus exists

Working principle of SMT (1 of 2)

- Very large data set of approved translations.
- Translation Model ➔ Frequency of phrases.
- More frequently a phrase is repeated ➔ More probable the target translation is correct.
- Probability model ➔ Target translation.

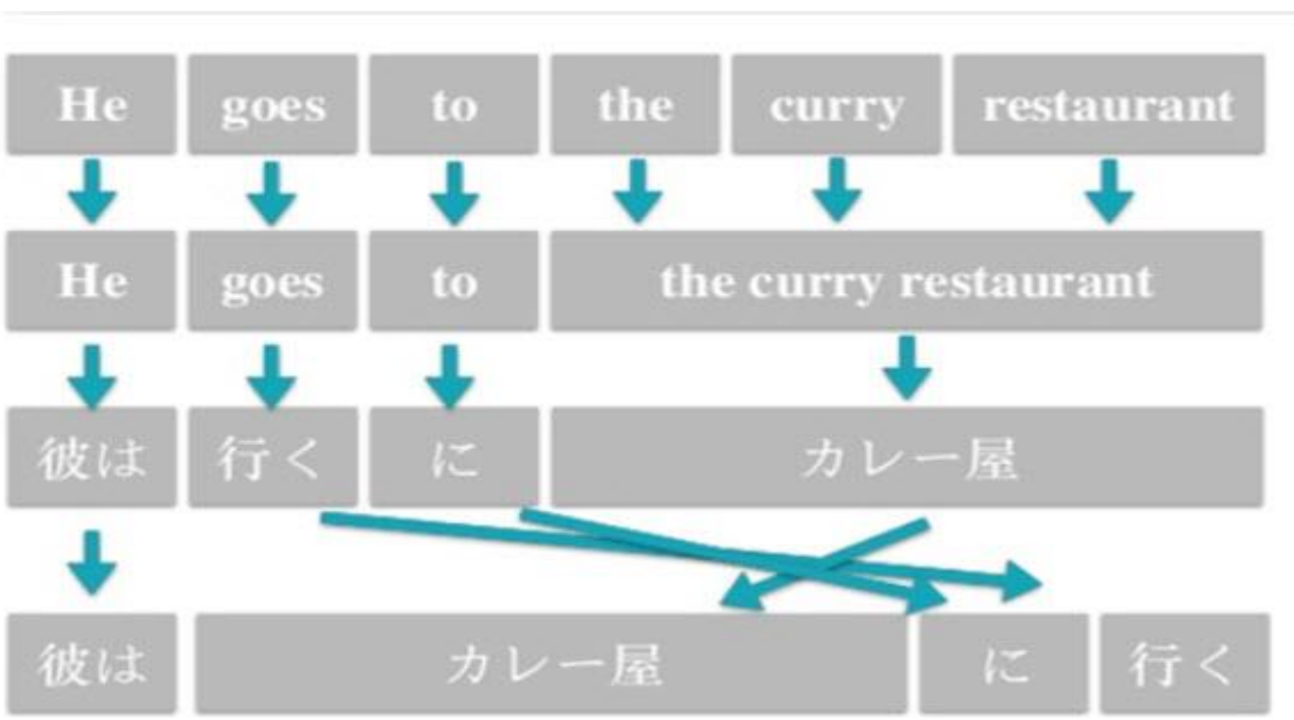


Figure: Working principles of SMT

Source: https://kantanmtblog.com/2019/04/02/a-short-introduction-to-the-statistical-machine-translation-model/

Working principle of SMT (2 of 2)

- Statistical machine translation.
- Decoding process.
- Speech to speech machine translation.

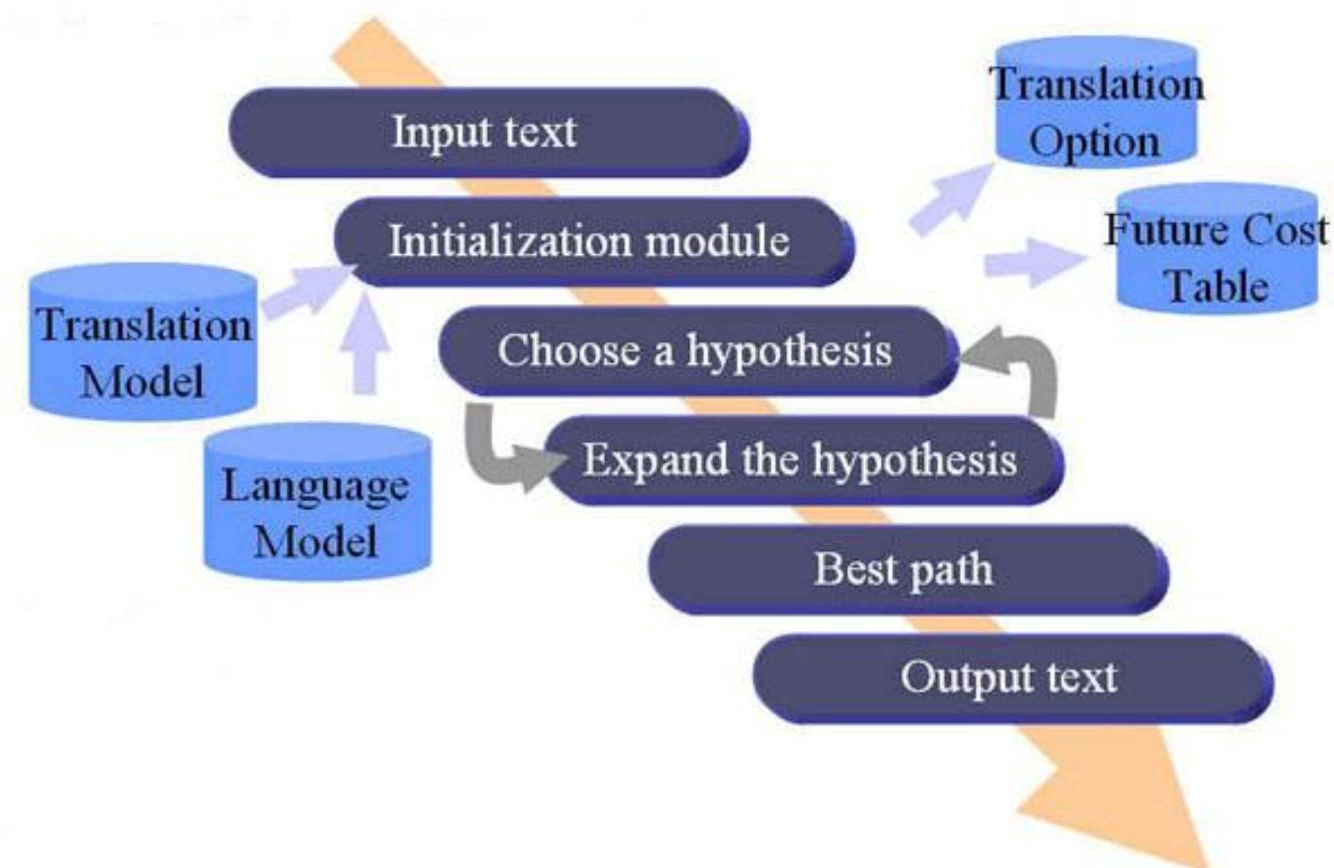


Figure: Working principles of SMT

Source: http://nlp.postech.ac.kr/research/previous_research/smt/

Challenges with statistical machine translation

- Sentence alignment:
  - Single sentences ➔ Translated into several sentences.
- Word alignment:
  - Words have no clear equivalent in the target language.
  - "John does not live here," ➔ "John wohnt hier nicht.“
- Statistical anomalies:
  - Override translations ➔ Proper nouns.
  - "I took the train to Berlin" = "I took the train to Paris“.
- Idioms:
  - Idioms may not translate "idiomatically".
  - "hear" ➔ "Bravo!" Parliament "Hear, Hear!" becomes "Bravo!“.
- Different word orders:
  - Word order in languages differ.
  - SVO or VSO languages.
- Out Of Vocabulary (OOV) words:
  - Different word forms as separate symbols without any relation.

Self evaluation: Exercise 13

- To continue with the training, after learning the concepts of Machine Translation and the Statistical methods in Natural Language Text Processing, it is time to write code to work with Tokenization and implement VITERBI algorithm. It is instructed to utilize the concepts of reading data from Treebank, Tokenization, Machine Translation with Keras library and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 13: Perform POS tagging implementing the Hidden Markov Model with Simple VITERBI algorithm and Rule Based VITERBI algorithm using NLTK Treebank. Split the Dataset for validation. Perform POS tagging on any text and compare the performance of the two representations of VITERBI algorithm.

Self evaluation: Exercise 14

- To continue with the training, after learning the concepts of Machine Translation and the Statistical methods in Natural Language Text Processing, it is time to write code to work with Tokenization and implement VITERBI algorithm. It is instructed to utilize the concepts of reading data from Treebank, Tokenization, Machine Translation with Keras library and perform the following activity.

- You are instructed to write the following activities using Python code.

- Exercise 14: Perform Machine Translation from one language to another. (German to English).

Checkpoint (1 of 2)

Multiple choice questions:

1. What is the main challenge/s of NLP?
   a) Handling ambiguity of sentences
   b) Handling tokenization
   c) Handling POS-tagging
   d) All the above

2. What is machine translation?
   a) Converts one human language to another
   b) Converts human language to machine language
   c) Converts any human language to English
   d) Converts Machine language to human language

3. What is morphological segmentation?
   a) Does discourse analysis
   b) Separate words into individual morphemes and identify the class of the morphemes
   c) Is an extension of propositional logic
   d) None of the above

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. What is the main challenge/s of NLP?
    - **a) Handling ambiguity of sentences**
    - b) Handling tokenization
    - c) Handling POS-tagging
    - d) All the above

2. What is machine translation?
    - **a) Converts one human language to another**
    - b) Converts human language to machine language
    - c) Converts any human language to English
    - d) Converts Machine language to human language

3. What is morphological segmentation?
    - a) Does discourse analysis
    - **b) Separate words into individual morphemes and identify the class of the morphemes**
    - c) Is an extension of propositional logic
    - d) None of the above

Checkpoint (2 of 2)

Fill in the blanks:

1. Many words have more than one meaning; we must select the meaning which makes the most sense in context. This can be resolved by ________________.
2. In a rule-based system, the form of procedural domain knowledge ______________.
3. Types are available in machine learning are _____.
4. ____________ are used to identify text based upon the same text as input.

True or False:

1. Speech segmentation is a subtask of speech recognition. True/False
2. Modern NLP algorithms are based on machine learning, especially statistical machine learning. True/False
3. Statistical machine translation uses algorithms for learning how to analyze the human translations. True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. Many words have more than one meaning; we must select the meaning which makes the most sense in context. This can be resolved by **Word sense disambiguation.**
2. In a rule-based system, the form of procedural domain knowledge **production rules.**
3. Types are available in machine learning are **3.**
4. **Variational auto encoders** are used to identify text based upon the same text as input.

True or False:

1. Speech segmentation is a subtask of speech recognition. **True**
2. Modern NLP algorithms are based on machine learning, especially statistical machine learning. **True**
3. Statistical machine translation uses algorithms for learning how to analyze the human translations. **True**

Question bank

Two mark questions:

1. What are multi word expressions?
2. What is cosine similarity?
3. What is WSD? Why is it needed?
4. What are parse trees?

Four mark questions:

1. Describe left corner parsing with examples.
2. How are multi word expressions classified?
3. Describe the methods of identifying text similarity.
4. What are the complications I word sense disambiguation?

Eight mark questions:

1. Write in detail about the concepts and process involved in statistical machine translation.
2. Discuss in detail about the various approaches to parsing.

Unit summary

Having completed this unit, you should be able to:

- Understand what is statistical parsing and the core concepts involved in it
- Learn about multiword expressions and how to handle them
- Understand the concepts of word similarity and the relativeness calculations done
- Gain knowledge on word sense disambiguation and why it is needed in NLP
- Gain an insight into modern speech recognition techniques with an idea of the forerunners in the field
- Understand what statistical machine translation means and the guidelines needed to perform SMT