UNIT 4 – NLP Applications

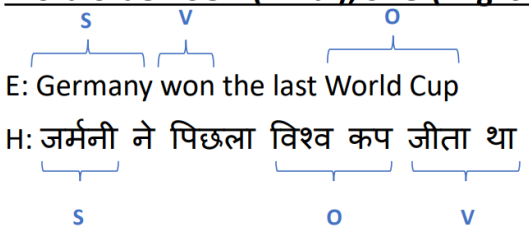
(Disclaimer: Many content is coped from various sources in internet and for academic purpose only)

Machine Translation

Machine translation is use of either rule-based or probabilistic machine learning approaches to translation of text or speech from one language to another.

Language divergence is the great diversity among the languages of the world. The central problem of MT is to bridge this language divergence.

Word order: SOV (Hindi), SVO (English)



Free (Hindi) vs rigid (English) word order

पिछला विश्व कप जर्मनी ने जीता था *(correct)*

The last World Cup Germany won *(grammatically incorrect)*
The last World Cup won Germany *(meaning changes)*

Why MT difficult?

Ambiguity:

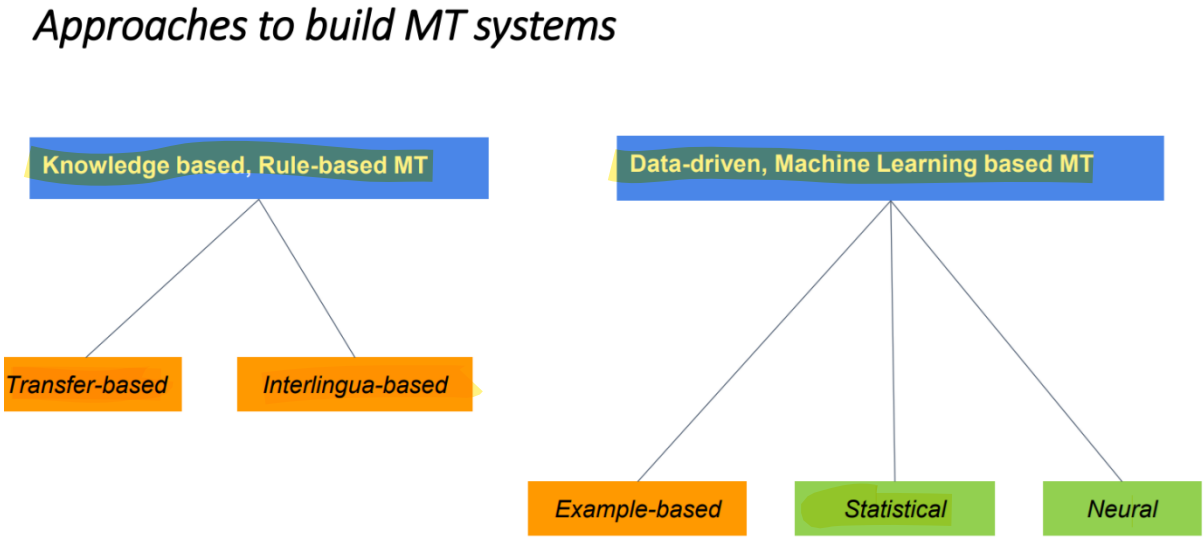
1. Same word multiple meaning: मंत्री (minister or chess piece)
2. Same meaning, multiple words जल, पानी, नीर (water)

Word Order:

1. Underlying deeper syntactic structure
2. Phrase structure grammar
3. Computationally intensive

Morphological Richness:

1. Identifying basic units/internal structure of words



The rule based machine translation paradigm includes

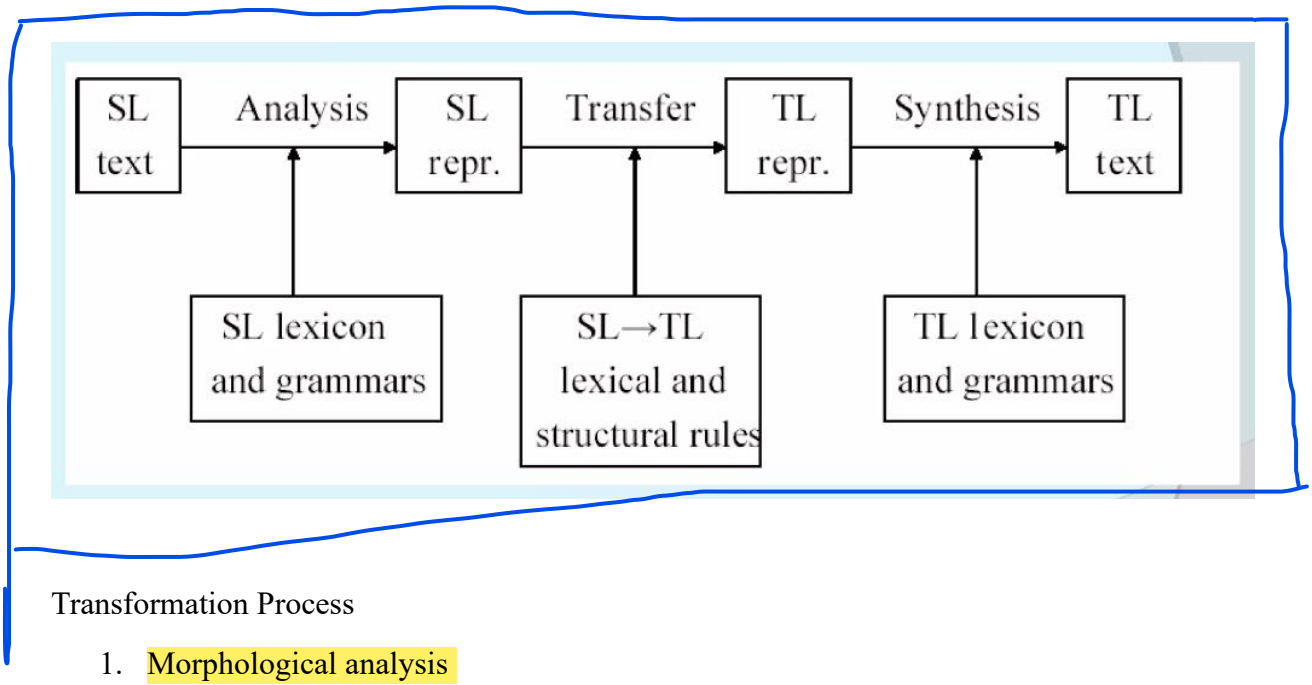
1. Transfer based machine translation
2. Interlingual machine translation and
3. Dictionary based machine translation

Transfer Based:

- If is necessary to have **an intermediate representation** that captures the "meaning" of original sentence in order to generate the correct translation
- In Interlingua-based MT this intermediate representation must be independent of the languages in question, whereas in transfer based, it has some dependence on the language pair involved.

Steps:

1. The original text is first analysed morphologically and syntactically, in order to obtain syntactic representation
2. This representation can then be refined to a more abstract level putting emphasis on the parts relevant for translation and ignoring other types of information
3. The transfer process then converts this final representation (still in the original language to a representation of the same level of abstraction in the target language
4. These two representations are referred to as "intermediate" representations
5. From the target language representation, the stages are then applied in reverse.



Transformation Process

1. Morphological analysis

Surface forms of the input text are classified as

i. To parts-of-speech (eg: noun, verb, etc.)
ii. Sub-category (number, gender, tense etc.)

2. Lexical categorisation

In any given text some words may have more than one meaning, causing ambiguity in analysis. Lexical categorisation look at the context of a word to try and determine the correct meaning in the context of the input

3. Lexical transfer

This is basically dictionary translation. The source language lemma (perhaps with sense information) is looked up in a bilingual dictionary and the translation is chosen

4. Structural transfer

While the previous stages deal with words, this stage deals with larger constituents

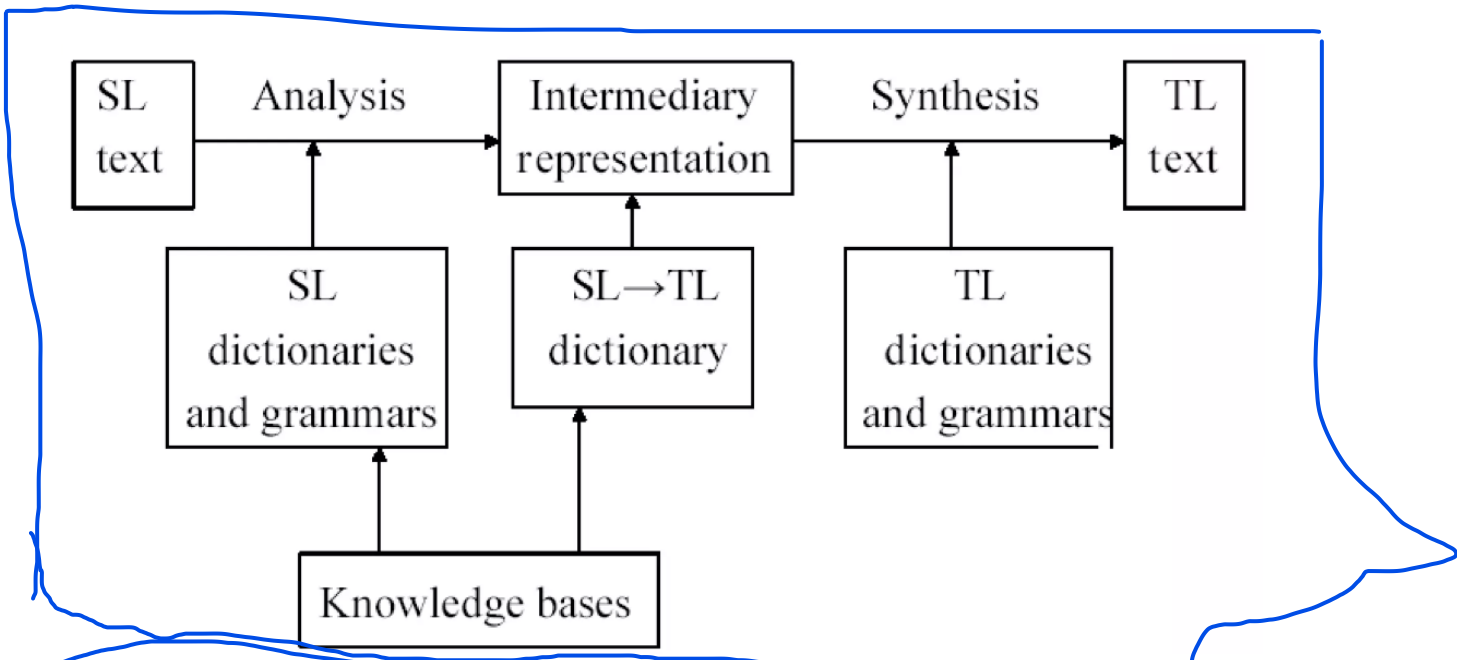
5. Morphological generation

From the output of the structural transfer stage, the target language surface forms are generated.

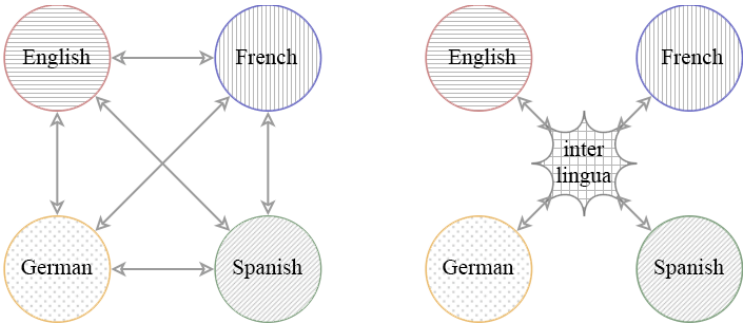
INTERLINGUAL MT

The source language, ie the text to be translated is transformed into an interlingua ie, an abstract language independent representation. The target language is then generated from the interlingua.

- In direct approach, words are translated directly without passing through an additional representation
- In transfer approach, the source language is transformed into an abstract, less language specific representation.



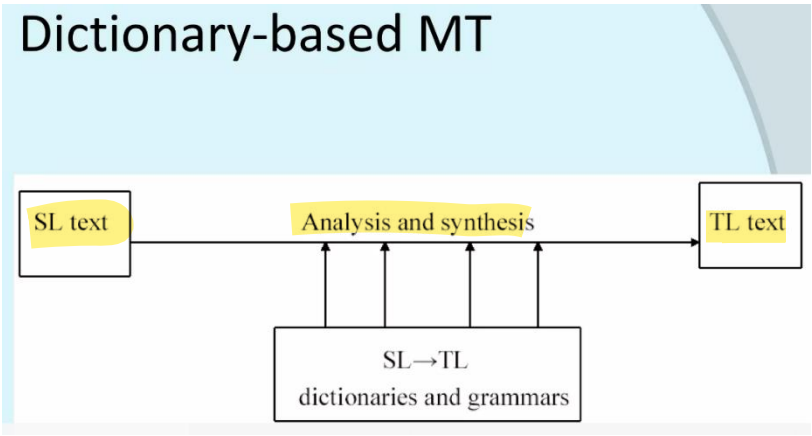
Transfer based Vs Interlingua

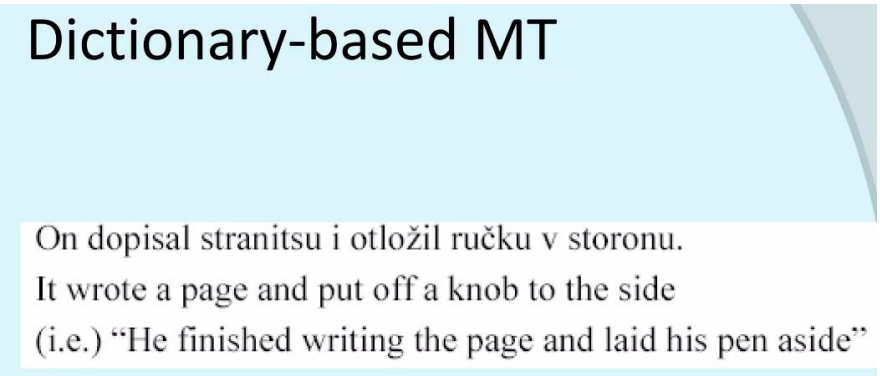


Dictionary Based MT

The words will be translated as a dictionary does – word by word, usually without much correlation of meaning between them. Dictionary lookups may be done with or without morphological analysis or lemmatisation.

Used to expedite manual translation, if the person carrying out is fluent in both language therefore capable of correcting syntax and grammer





Example Based MT

Characterised by its use of bilingual corpus with parallel texts as its main knowledge base. It is essentially a translation by analogy and can be viewed as an implementation of case based reasoning approach of machine learning

Bilingual parallel corpora example

English	Japanese
How much is that red umbrella?	Ano akai kasa wa ikura desu ka.
How much is that small camera?	Ano chiisai kamera wa ikura desu ka.

Statistical MT

The idea behind statistical machine translation comes from information theory. A document is translated according to the probability distribution P(e|f) that a string e in the target language (for example, English) is the translation of a string f in the source language (for example, French)

The problem of modelling the probability distribution p(e|f) has been approached in a number of ways. One intuitive approach is to apply Bayes Theorem.

$P(e|f) = p(f|e)p(e)$

Where the translation model p(f|e) is the probability that the source string is the translation of the target string, and the language model p(e) is the probability of seeing that target language string.

Finding the best translation is done by picking up the one that gives the highest probability

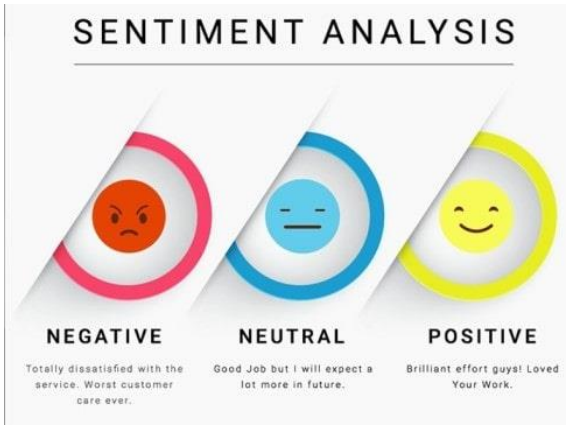
$$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e)$$

Sentiment Analysis

Sentiment analysis is a process of analyzing a piece of text and identifying opinions and judgments. This procedure shows you whether the sentiment of a piece of text is positive, negative, or neutral.

Natural Language Processing (NLP) and machine learning are used in unison to score the sentiments for categories, topics, or entities in a phrase. Along with ML and NLP, other kinds of data analysis techniques are also used to process and analyze the raw text and pull objective quantitative results from it.

Sentiment Analysis is also known as opinion mining or emotion AI. It can be used to transform unstructured text from social media, news, forums, blogs, etc., into structured data.



The purpose of sentiment analysis is to allow companies to understand the market's sentiments regarding their brands or products. It is rather important because it enables brands to understand how their customers and prospective customers feel about them.

Understanding customer sentiments allows businesses to make higher quality and more informed decisions.

When humans perform sentiment analysis, the results tend to be very subjective. It depends on people's personal experiences, their thoughts, beliefs, and many other factors. When you use an automated sentiment analysis system, your organization gets to to apply the same, uniform criteria to all your data, thus improving accuracy, creating consistency, and helping you gain better results.

Sentiment analysis helps you avoid PR nightmares. If someone told you that all publicity is good publicity, they're either lying or they don't know better. Bad publicity is certainly a thing, and you want to take care of it sooner rather than later. Not making use of sentiment analysis could result in your brand getting damaged substantially.

Applications to review related websites

- User reviews from topics, services, movies, products related websites

Application as a sub component technology

- Augmentation to recommendation system
- Detecting heating language in emails
- Detecting sensentive content webpages
- Sentiment oriented question answering system

Applications in business and government intelligence

- Finding customers attitude and trends
- Election results prediction

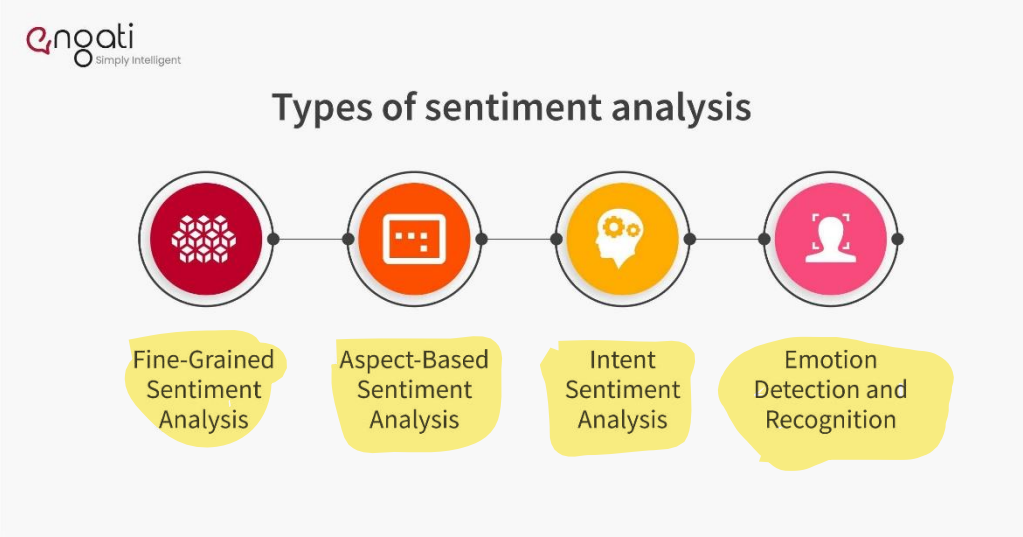
Applications across different domains

- Applications in politics, eRulemaking, and sociology

Other applications

- Prediction of sales performance
- Public opinion polls
- Box office revenues for movies
- Stock market prediction

Types



Fine-Grained Sentiment Analysis

Most sentiment analysis systems employ fine-grained sentiment analysis. It concentrates on the polarity of an opinion. Rather than rating sentiments to just be positive, negative, or neutral, it segments them further into very positive, somewhat positive, neutral, somewhat negative, and very negative.

It can pull this data from product reviews, surveys, etc.

The polarity detected can be tied to emotions as well.

Very positive = love

Somewhat positive = acceptance

Neutral = indifference

Somewhat negative = anxiety

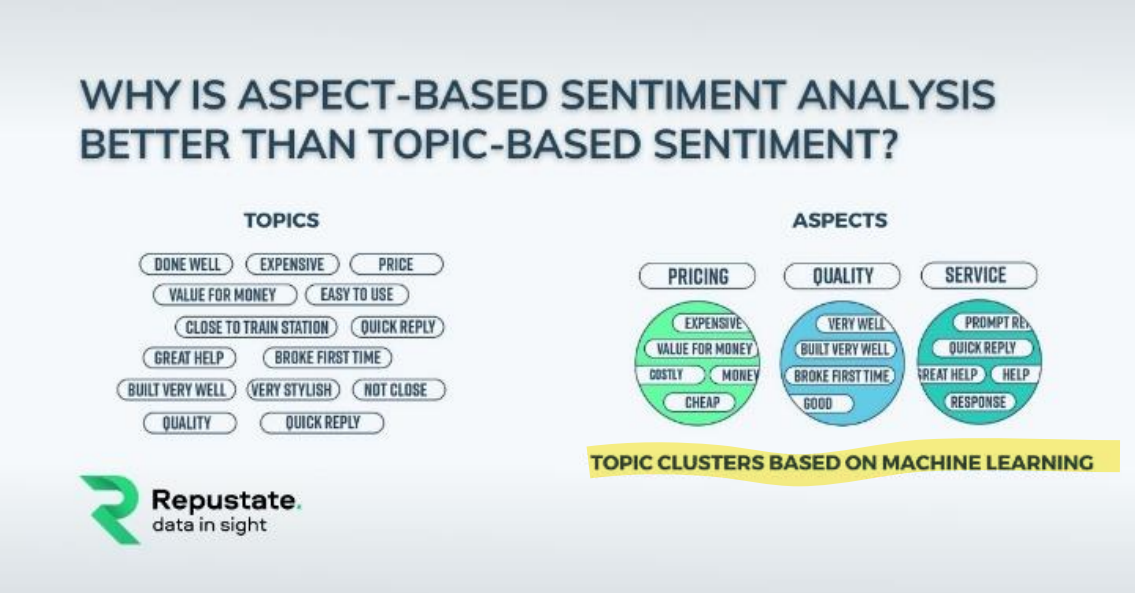
Very negative = anger

Aspect-Based Sentiment Analysis

Instead of just looking at polarity, aspect-based sentiment analysis allows you to consider the sentiment regarding specific aspects, features, or elements of your product or service.

It helps you understand how your customers feel about specific attributes of your offerings.

It is essentially a text analysis method that classifies sentiments according to aspects and detects the sentiment that is attributed to each one. It also allows organizations to automatically analyze enormous amounts of data in detail, thus saving money and time and allowing teams to focus on more important tasks.



Intent Sentiment Analysis

This is about understanding the intent that lies behind a message. Is it an opinion? Does it show appreciation? Is it a complaint, a question, or a suggestion?

However, since it is difficult to do this without context, this is still a concept and is not practically used yet. It involves making use of Long Short Term Memory (LSTM) algorithms to categorize text into various LSTMs model sentences as chain of forget-remember decisions based on context

Emotion Detection and Recognition

This relies more on algorithms and uses lexicons and machine learning to understand customer emotions. But it is not just limited to simply showing you the emotion, it helps you understand why customers feel how they do.

The majority of emotion detection systems make use of lexicons (lists of words along with the emotions that they convey) or complex machine learning algorithms.

But just using lexicons might not always be the best idea. This is because people tend to express emotions in various ways. Words that typically convey negative emotions could also be used to expression positive ones. For example, In the phrase 'It makes me sick', the word 'sick' conveys a negative emotion, but in the phrase 'This is sick! I love it', the word 'sick' conveys a positive emotion.

CHATBOT

A chatbot is a software program for simulating intelligent conversations with humans using rules or artificial intelligence. A computer program that can comprehend human language and communicate with a user via a website or messaging app is known as a chatbot (conversational interface, AI agent). **Chatbots** are conversational technologies that effectively carry out repetitive activities. They are well-liked by people because they facilitate the speedy completion of those tasks, freeing them up to concentrate on more complex, strategic, and interesting duties that call for human qualities that are unmatched by computers.

Types of Chatbots

1. Menu or Button-Based Chatbot

The simplest type of chatbots are menu-based or button-based chatbot, in which users can communicate with them by selecting the button from a scripted menu that most closely matches their requirements. The user-friendly chatbot may present a new set of possibilities based on their clicks, which they can proceed to select until they arrive at the most appropriate and targeted option. These chatbots function essentially like a decision tree.

2. Rule-Based Chatbot

Expanding on the basic decision tree capability of the menu-based chatbot, the rules-based chatbot utilises conditional if/then logic to create automated conversation flows. Rule-based bots function similarly to interactive FAQs, with the conversation designer programming preset question-and-answer combinations into the bot so it can comprehend user input and provide relevant responses.

3. AI Powered Chatbots

AI chatbots can comprehend user questions regardless of how they are expressed, however the conversational flow of rules-based chatbots only allows predefined questions and answer possibilities. The AI bot's natural language understanding (NLU) and artificial intelligence (AI) skills enable it to promptly identify any pertinent contextual information that the user shares, facilitating a more conversational and seamless exchange of ideas. The AI-powered chatbot may pose clarifying questions when it is unclear what a user is requesting and discovers multiple actions that could satisfy the request.

4. Voice Chatbots

A Voice bot is a type of artificial intelligence (AI) software that can converse with incoming calls in contact centres. Using natural language processing (NLP) and machine learning, it records, decodes, and interprets voice input from users and answers intelligibly. Another conversational technology that lets users engage with the bot by speaking to it instead of typing is called voice chatbot. Certain voice chatbots are more basic than others.

Applications of Chatbots

- **Customer Service and Support:** Chatbots are substantially utilized in customer support to provide help and assistance to customers. They can answer frequently asked questions, troubleshoot problems, and guide users via self-carrier options, thereby reducing wait times and enhancing overall customer experience.
- **E-trade and Retail:** In the e-trade quarter, chatbots are employed to assist buyers with product hints, answer inquiries about product capabilities, pricing, and availability, and facilitate seamless transactions.
- **Financial Services:** Chatbots can help users manipulate their finances, music prices, set budgets, and even offer investment pointers primarily based on person desires and chance profiles.
- **Healthcare :** In healthcare, chatbots are applied for a number of functions, including appointment scheduling, medicine reminders, symptom evaluation, and patient schooling. They can triage affected person inquiries, offer fundamental clinical advice, and offer assist for intellectual fitness and well-being.
- **Education and Training:** Chatbots are widely used in schooling and education settings to supply customized mastering education, provide tutoring and academic courses.

Benefits of Chatbots

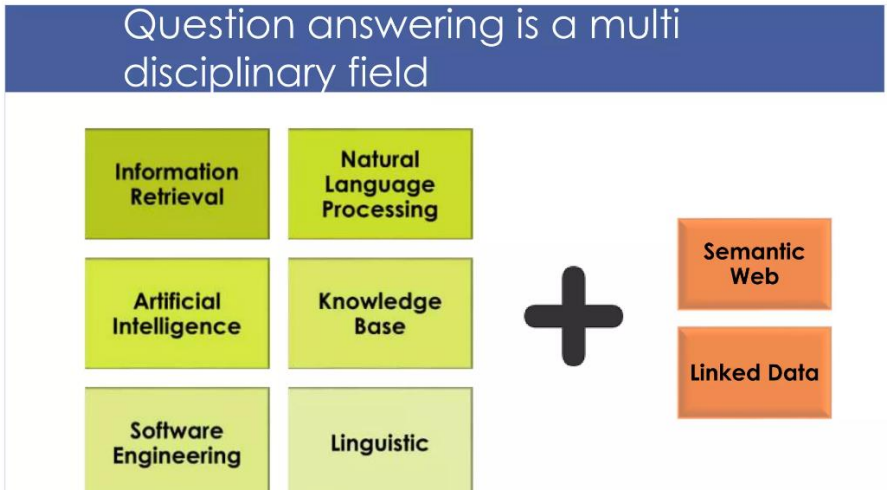
- **24/7 Availability:** Chatbots can provide contimous support, improving customer service accessibility.
- **Scalability:** They can take care of a couple of conversations concurrently, scaling effortlessly to accommodate growing demand.
- **Cost Savings:** Automation reduces the need for human agents, resulting in fee savings for agencies.
- **Instant Responses:** Chatbots deliver immediate responses, improving reaction times and person delight.
- **Consistency**: They provide regular and standardized responses, ensuring uniformity in customer interactions.
- **Data Collection:** Chatbots can acquire precious consumer statistics and insights, enabling personalized stories and centered advertising and marketing.
- **Integration:** Chatbots can combine with present systems and structures, enhancing workflow efficiency.

Limitations of Chatbots

- **Lack of Understanding:** Chatbots may also struggle to recognize complicated queries or nuances in language, which leads to misunderstandings.
- **Limited Scope:** They are powerful as their programmed abilties, restricting their ability to handle unforeseen data.
- **Impersonal Interactions:** Some customers may also decide on human interplay over interacting with a system, leading to dissatisfaction.
- **Technical Issues:** Chatbots are prone to technical system faults and errors, that can disrupt consumer reports.
- **Dependency:** Overreliance on chatbots may result in reduced human interaction and lack of interpersonal connections.
- **Initial Investment:** Developing and imposing chatbots requires preliminary investment in phrases of time, resources, and technical kmowledge.
- **Maintenance:** Chatbots require ongoing protection and updates to stay powerful and relevant through the years.

Question Answering system

Question answering is a computer science discipline within the fields of information retrieval and natural language processing that is concerned with building systems that automatically answer questions that are posed by humans in a natural language.



Difference to Information Retrieval

- Information Retrieval is a **query driven** approach for accessing information.
  - System returns a list of documents.
  - It is responsibility of user to navigate on the retrieved documents and find its own information need.
- Question Answering is an **answer driven** approach for accessing information.
  - User asks its question in natural language (i.e. phrase-based, full sentence or even keyword based) queries.
  - System returns the list of short answers.
  - More complex functionality.

Natural language queries are classified into different categories

- **Factoid queries**: WH questions like when, who, where.
- **Yes/ No queries**: Is Berlin capital of Germany?
- **Definition queries**: what is leukemia?
- **Cause/consequence queries**: How, Why, What. what are the consequences of the Iraq war ?

Types of Questions in Modern Systems

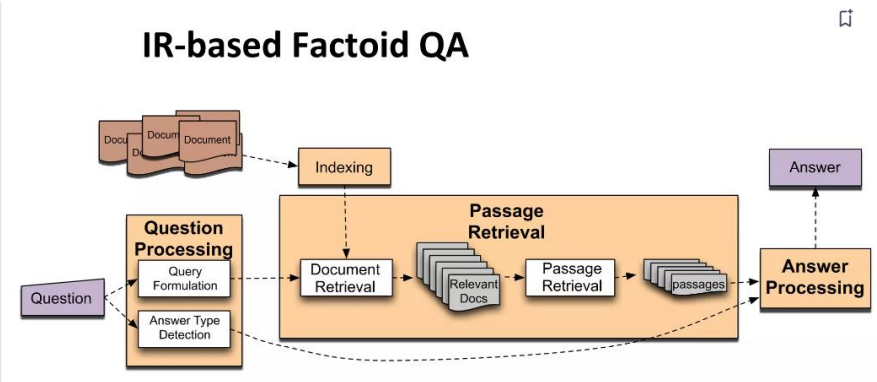
- Factoid questions
  - *Who wrote "The Universal Declaration of Human Rights"?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
  - *What do scholars think about Jefferson's position on dealing with pirates?*

Open Datasets available for Question Answering

1. *Stanford Question Answering Dataset (SQuAD)*
2. WikiQA dataset [3], is a publicly available set of question and answer pairs, collected and annotated for research on open-domain question answering.
3. The TREC-QA dataset contains questions and answer patterns, as well as a pool of documents returned by participating teams.
4. NewsQA dataset [4] is to help the research community build algorithms that are capable of answering questions requiring human-level comprehension and reasoning skills.

Types of Question Answering

1. IR-based Factoid Question Answering



1. Question processing
   - Detect question type, answer type, focus, relations
   - Formulate queries to send to a search engine

2. Passage retrieval
    - Retrieve ranked documents
    - Break into suitable passages and rerank
3. Answer Processing

- Extract candidate answers
- Rank candidates – using evidence from the text and external sources

Knowledge Based Approaches (Siri)

Build a semantic representation of the query

- Times, dates, locations, entities, numeric quantities

Map from this semantics to query structured data or resources

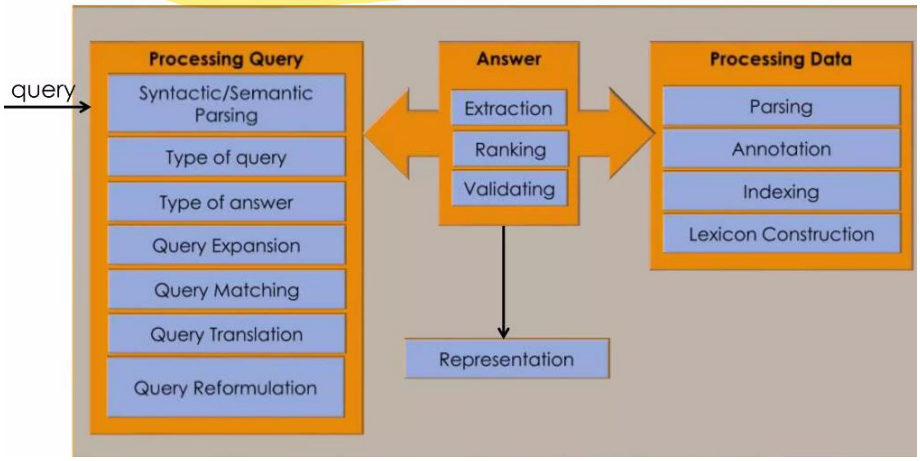
- Geospatial databases
- Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
- Restaurant review sources and reservation services
- Scientific databases

Types of QA systems

- Open-domain: domain independent QA systems can answer any query from any corpus
    - \+ covers wide range of queries
    - \- low accuracy
- Closed-domain: domain specific QA systems are limited to specific domains
    - \+ High accuracy
    - \- limited coverage over the possible queries
    - \- Needs domain expert

Some concepts in QA

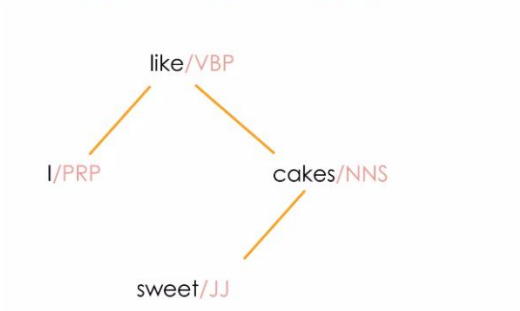
Core of a QA system



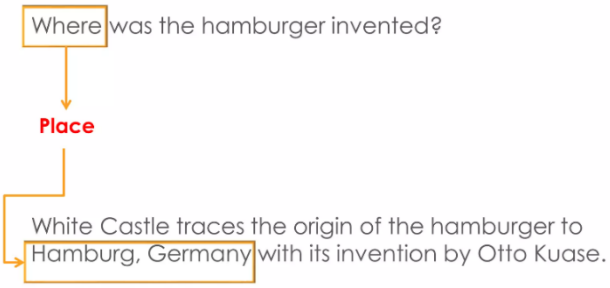
Syntactic Parsing: Part-of-speech Tagging

I like sweet cakes

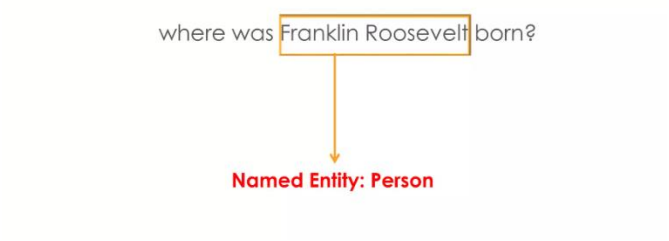
I/PRP like/VBP sweet/JJ cakes/NNS



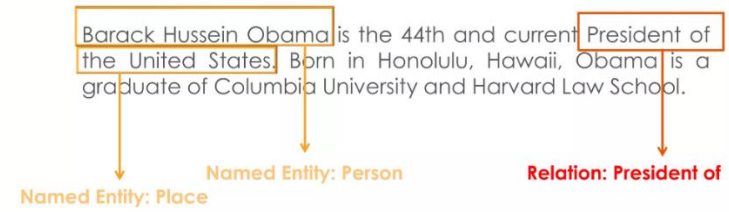
Type of Answer



Named Entity Recognition on Query



Relation Extraction



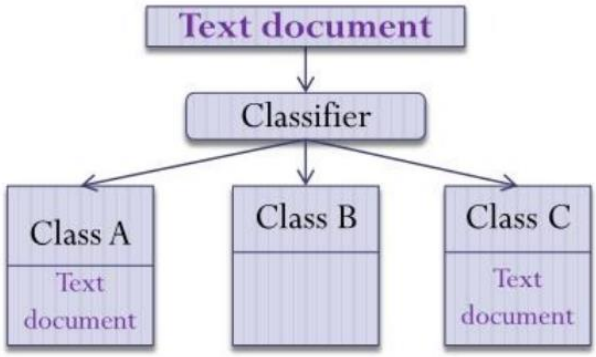
Watson Project

- Watson is a computer which is capable of answering question issued in natural language.
- Questions come from quiz show called Jeopardy.
- The software of this project is called DeepQA project.
- In 2011, Watson won the former winners of quiz show Jeopardy.

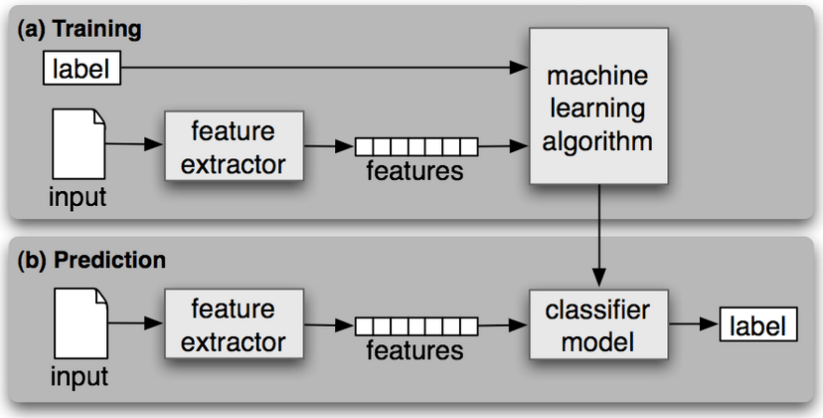
Decomposition example

Text Classification

- Text classification is a task that involves categorizing or classifying text documents into predefined categories or classes
- It plays a crucial role in various applications, including sentiment analysis, spam detection and topic classification



Text classification workflow



A. Data preprocessing

- Preprocess the data by cleaning, tokenizing, and normalizing the text.
- Techniques like removing stopwords, stemming or lemmatisation may be applied.

B. Feature Extraction

- Represent the text documents as numerical feature vectors that machine learning algorithms can process
- Techniques like bag-of-words, TF-IDF, or word embeddings (eg: Word2Vec or Glove) are commonly used for feature extraction

C. Training data preparation

- Prepare a labelled training dataset with text documents and their corresponding class labels
- Ensure a balanced distribution of classes and a sufficient number of training samples for each class.

D. Model training

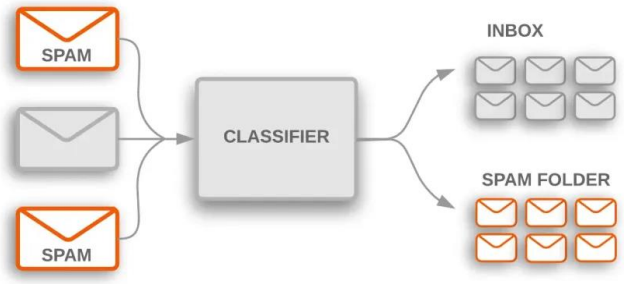
- Train a text classification model using machine learning algorithms or deep learning architectures.

- Algorithms like Naïve Bayes, Support Vector Machines (SVM), Random forest, Neural Networks can be employed

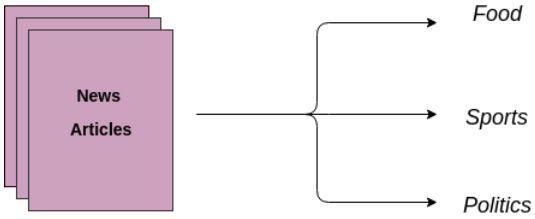
E. Model Evaluation

- Assess the performance of the trained model using evaluation metrics such as accuracy, precision, recall or F1 score.
- Use the techniques like cross-validation or holdout validation to evaluate the model's generalization ability

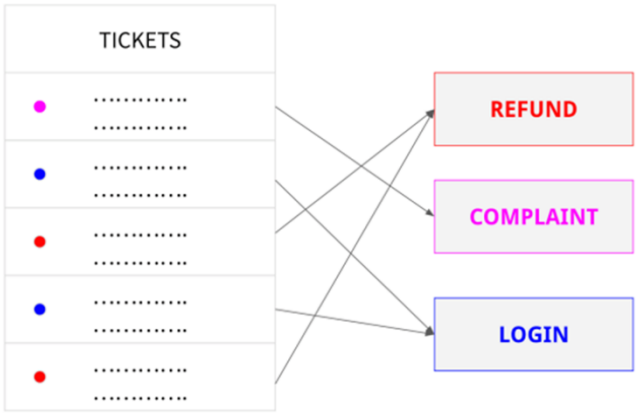
Spam classification



Classifying news articles and blogs



Categorize customer support requests



Types of Text Classification Systems

Rule-based text classification

Rule-based techniques use a set of manually constructed language rules to categorize text into categories or groups. These rules tell the system to classify text into a particular category based on the content of a text by using semantically relevant textual elements. An antecedent or pattern and a projected category make up each rule.

For example, imagine you have tons of new articles, and your goal is to assign them to relevant categories such as Sports, Politics, Economy, etc.

With a rule-based classification system, you will do a human review of a couple of documents to come up with linguistic rules like this one:

- If the document contains words such as *money*, *dollar*, *GDP*, or *inflation*, it belongs to the Politics group (class).

Rule-based systems can be refined over time and are understandable to humans. However, there are certain drawbacks to this strategy.

These systems, to begin with, demand in-depth expertise in the field. They take a lot of time since creating rules for a complicated system can be difficult and frequently necessitates extensive study and testing.

Given that adding new rules can alter the outcomes of the pre-existing rules, rule-based systems are also challenging to maintain and do not scale effectively.

Machine learning-based text classification

Machine learning-based text classification is a supervised machine learning problem. It learns the mapping of input data (raw text) with the labels (also known as target variables). This is similar to non-text classification problems where we train a supervised classification algorithm on a tabular dataset to predict a class, with the exception that in text classification, the input data is raw text instead of numeric features.

Like any other supervised machine learning, text classification machine learning has two phases; training and prediction.

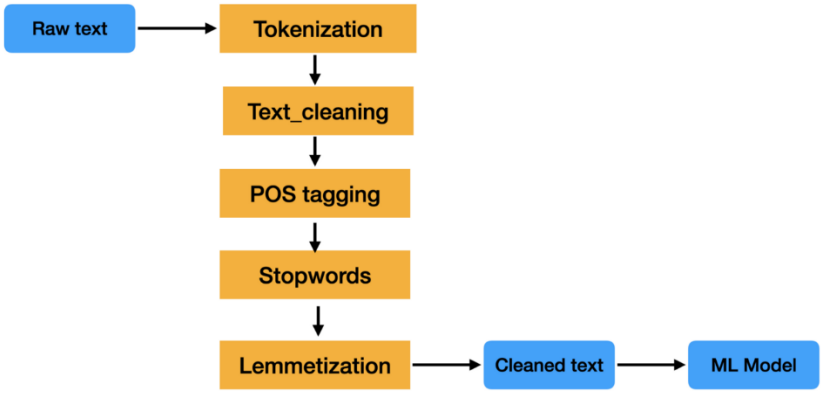
Training phase

A supervised machine learning algorithm is trained on the input-labeled dataset during the training phase. At the end of this process, we get a trained model that we can use to obtain predictions (labels) on new and unseen data.

Prediction phase

Once a machine learning model is trained, it can be used to predict labels on new and unseen data. This is usually done by deploying the best model from an earlier phase as an API on the server.

Text Preprocessing Pipeline



Feature Extraction

The two most common methods for extracting feature from text or in other words converting text data (strings) into numeric features so machine learning model can be trained are: Bag of Words (a.k.a CountVectorizer) and Tf-IDF.

(Learn gender classification, document classification, pos classification from NLTK)

Spell checking

Who cares about spelling?
Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Detection vs. Correction
There are two distinct tasks:error detection = simply find the misspelled wordserror correction = correct the misspelled wordse.g., It might be easy to tell that ater is a misspelled word, but what is the correct word? water? later? after?So, what causes errors?

Keyboard mistyping Space bar issues Keyboard proximity
run-on errors = two separate words become one

e.g., the fuzz becomes thefuzz

split errors = one word becomes two separate words

e.g., equalization becomes equali zation

Keyboard proximity

e.g., Jack becomes Hack since h, j are next to each other on a typical American keyboard

Physical similarity

similarity of shape, e.g., mistaking two physically similar letters when typing up something handwritten e.g., tight for fight

Phonetic errors

phonetic errors = errors based on the sounds of a language (not necessarily on the letters)

homophones = two words which sound the same

e.g., red/ read (past tense), cite/ site/ sight, they're/ their/ there

Spoonerisms = switching two letters/sounds around

e.g., It's a ravy(vary) grain with biscuit wheels.

letter substitution = replacing a letter (or sequence of letters) with a similar-sounding one

e.g., John kracked his nuckles. instead of John cracked his knuckles

e.g., I study sikologee.

Knowledge problems

And then there are simply cases of not knowing how to spell:

not knowing a word and guessing its spelling (can be phonetic)

e.g., scientist

not knowing a rule and guessing it

e.g., Do we double a consonant for ing words? Jog -> joging ?

What makes spelling correction difficult?

Tokenization: What is a word?

Definition is difficult with contractions, multi-token words, hyphens, abbreviations

Inflection: How are some words related?

How do we store rules and exceptions?

Productivity of language: How many words are there?

Words entering and exiting the lexicon

How we handle these issues determines how we build a dictionary.

Techniques used for spell checking

Non-word error detection

Isolated-word error correction

Context-dependent word error detection and correction

Real-word errors

grammar correction.

The exact techniques used will differ depending on if we are looking for spelling errors in human typing or with optical character recognition (OCR)

Non-word error detection
non-word error detection is essentially the same thing as word recognition = splitting up "words" into true words and non-words.

How is non-word error detection done?

Using a dictionary: Most common way to find non-word errors

N-gram analysis: used with OCR more than typing

Dictionaries Intuition: Two aspects:
Have a complete list of words and check the input words against this list.

If it's not in the dictionary, it's not a word.

Two aspects:

Dictionary construction = build the dictionary (what do you put in it?)

Dictionary lookup = lookup a potential word in the dictionary (how do you do this quickly?)

Dictionary construction
Do we include inflected words? i.e., words with prefixes and suffixes already attached

Pro: lookup can be faster

Con: takes much more space, doesn't account for new formations

Want the dictionary to have only the word relevant for the user -> domain-specificity

e.g., For most people memoize is a misspelled word, but in computer science this is a technical term and spelled correctly.

Foreign words, hyphenations, derived words, proper nouns, and new words will always be problems for dictionaries since we cannot predict these words until humans have made them words.

Dictionary should probably be dialectally consistent.

e.g., include only color or colour but not both

Dictionary lookup Several issues arise when trying to look up a word:
Have to make lookup fast by using efficient lookup techniques, such as a hash table (cf. the indices we discussed under the searching topic)

Have to strip off prefixes and suffixes if the word isn't an entry by itself.

N-gram analysis

Instead of storing possible words in a language, we store possible n-grams of letters

Check the input against the n-grams in the database

Can also store n-grams for certain word positions—e.g., mm is a possible bigram in English, but not as a word beginning

A fast and simple technique, but most typing errors are still valid n-grams, so this is more useful for OCR

Knowledge about typical errors

Word length effects: most misspellings are within two characters in length of originalWhen searching for the correct spelling, we do not usually need to look at words with greater length differencesFirst-position error effects: the first letter of a word is rarely erroneousWhen searching for the correct spelling, the process is sped up by being able to look only at words with the same first letter.

Isolated-word error correction methods

Many different methods are used; we will briefly look at four methods:

rule-based methods

similarity key techniques

minimum edit distance

probabilistic methods.

The methods play a role in one of the three basic steps:1. Detection of an error (discussed above) 2. Generation of candidate corrections 3. Ranking of candidate corrections

Rule-based methods

One can generate correct spellings by writing rules:

Common misspelling rewritten as correct word:

e.g., hte -> the

Rules

based on inflections:

e.g., V+C+ing -> V+CC+ing (where V = vowel and C = consonant)

based on other common spelling errors (such as keyboard effects or common transpositions):

e.g., Cie -> Cei

Similarity key techniques

Problem: How can we find a list of possible corrections?

Solution: Store words in different boxes in a way that puts the similar words together.

Example:

1. Start by storing words by their first letter (first letter effect),

e.g., punc starts with the code P.

2. Then assign numbers to each letter

b, f, p, v →1

c, g, j, k, q, s, x, z →2

d, t → 3

l → 4

m, n →5

r →6

e.g., punc -> P052

3. Then throw out all zeros and repeated letters,

e.g., P052 -> P52.

4. Look for real words within the same box,

e.g., punk is also in the P52 box.

How is a mistyped word related to the intended?

Types of errors

insertion = a letter is added to a word

deletion = a letter is deleted from a word

substitution = a letter is put in place of another one

transposition = two adjacent letters are switched

Note that the first two alter the length of the word, whereas the second two maintain the same length.

General properties

single-error misspellings = only one instance of an error

multi-error misspellings = multiple instances of errors (harder to identify)

Probabilistic methods

Two main probabilities are taken into account:

transition probabilities = probability (chance) of going from one letter to the next.

. e.g., What is the chance that a will follow p in English? That u will follow q?

confusion probabilities = probability of one letter being mistaken (substituted) for another (can be derived from a confusion matrix)

e.g., What is the chance that q is confused with p?

Useful to combine probabilistic techniques with dictionary methods

Confusion probabilities

For the various reasons discussed above (keyboard layout, phonetic similarity, etc.) people type other letters than the ones they intended.

It is impossible to fully investigate all possible error causes and how they interact, but we can learn from watching how often people make errors and where.

One way of doing so is to build a confusion matrix = a table indicating how often one letter is mistyped for another (this is a substitution matrix)

Confusion Matrix

del[x,y]: number of times in the training set 'xy' is the correct where 'x' typed

ins[x,y]: number of times in the training set x is correct where 'xy' typed

sub[x,y]: number of times that x was typed as y

trans[x,y]: number of times that xy was typed yx

The Noisy Channel Model

We can view the setup like this:

SOURCE: word -> NOISY CHANNEL -> noisy word

We need to decode the noisy word to figure out what the original was

The noisy channel model has been very popular in speech recognition, among other fields

Noisy word: O = observation (incorrect spelling)

To guess at the original word, we want to find the word (w) which maximizes: P(w|O), i.e., the probability of w, given that O has been seen

Conditional probability

p(x| y) is the probability of x given y

Let's say that yogurt appears 20 times in a text of 10,000 words

p(yogurt) = 20/10, 000 = 0.002

Now, let's say frozen appears 50 times in the text, and yogurt appears 10 times after it

p(yogurt| frozen) = 10/50 = 0.20

Bayes' Law

P(w|O) is very hard (impossible?) to calculate directly, but the following can be estimated easily from larger texts (more on that soon):

P(O|w) = the probability of the observed misspelling given the correct word

P(w) = the probability of the word occurring anywhere in the text

Bayes' Law allows us to calculate p(x| y) in terms of p(y| x):

(1) P(x| y) =P(y| x) P(x) / P(y)

(2) P(w|O) = P(O|w) P(w) / P(O)

Bayes' Law for Spelling

We can ignore the denominator because P(O) will be the same for every correction

(3) a. P(w1|O) ≈ P(O|w1)P(w1) b. P(w2|O) ≈ P(O|w2)P(w2)

1. List "all" possible candidate corrections, i.e., all words with one insertion, deletion, substitution, or transposition

2. Rank them by their probabilities

Minimum edit distance

In order to rank possible spelling corrections more robustly, we can calculate the minimum edit distance = minimum number of operations it would take to convert one word into another.

For example, we can take the following five steps to convert junk to haiku:

1. junk -> juk (deletion)

2. juk -> huk (substitution)

3. huk -> hku (transposition)

4. hku -> hiku (insertion)

5. hiku -> haiku (insertion)

But is this the minimal number of steps needed?

Market Intelligence

What is Market Intelligence? (youtube.com)