

NLP_UNIT-5

SHORT ANSWERS

[1] Anaphora

Ans: Reference in a text to an entity that has been previously introduced into the anaphora discourse is called anaphora, and the referring expression used is said to be an anaphor, or anaphoric.

[2] Co-reference resolution

Ans: Coreference resolution resolution is the task of determining whether two mentions corefer, by which we mean they refer to the same entity in the discourse model (the same discourse entity).

- The set of corefering expressions is often called a coreference chain or a cluster.

[3] Discourse Parsing

Ans: Discourse parsing is an important research area in natural language processing (NLP), which aims to parse the discourse structure of coherent sentences .

Given a sequence of sentences, how can we automatically determine the coherence relations between them? This task is often called discourse parsing (even though discourse parsing for PDTB we are only assigning labels to leaf spans and not building a full parse tree as we do for RST).

[4] Disambiguation/WSD

Ans: In natural language processing, word sense disambiguation (WSD) is the problem of determining which "sense" (meaning) of a word is activated by the use of the word in a particular context, a process which appears to be largely unconscious in people.

[5] Machine Translation

Ans: Machine Translation (MT) is the task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language.

[6] Applications of Nlp

Ans: Natural Language Processing is a cross among many different fields such as:

- Sentiment Analysis.
- Text Classification.
- Chatbots & Virtual Assistants.
- Text Extraction.
- Machine Translation.
- Text Summarization.
- Market Intelligence.
- Auto-Correct.

[7] Text Coherence

Ans: Coherence is a key property of any well-organized text. It evaluates the degree of logical consistency for text and can help document a set of sentences into a logically consistent order, which is at the core of many text-synthesis tasks such as text generation and multi- document summarization.

[8] Sentiment Analysis or Opinion mining

Ans: Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral.

Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

[9] Semantic role labelling

Ans: In natural language processing, semantic role labeling (also called shallow semantic parsing or slot-filling) is the process that

assigns labels to words or phrases in a sentence that indicates their semantic role in the sentence, such as that of an agent, goal, or result. It serves to find the meaning of the sentence.

[10] Selection Restriction

Ans: Selectional restrictions place semantic constraints on arguments and account for the implausibility of sentences such as Colorless green ideas slept furiously.

- They have been used in natural language understanding for disambiguation and pronoun resolution.
- A selectional restriction is a restriction concerning the relation between a predicate (or predicate term) and its argument(s) (argument terms). It concerns inherent properties of the relevant referents.

LONG ANSWERS

[1] 10 Applications of nlp in detail.

Ans:

Email filters:

Email filters are one of the most basic and initial applications of NLP online. It started out with spam filters, uncovering certain words or phrases that signal a spam message. But filtering has upgraded, just like early adaptations of NLP.

Smart assistants:

Smart assistants like Apple's Siri and Amazon's Alexa recognize patterns in speech thanks to voice recognition, then infer meaning and provide a useful response.

Search results:

Search engines use NLP to surface relevant results based on similar search behaviors or user intent so the average person finds what they need without being a search-term wizard.

Predictive text:

Things like autocorrect, autocomplete, and predictive text are so commonplace on our smartphones that we take them for granted. Autocomplete and predictive text are similar to search engines in that they predict things to say based on what you type, finishing the word or suggesting a relevant one.

Language translation:

One of the tell-tale signs of cheating on your Spanish homework is that grammatically, it's a mess. Many languages don't allow for straight translation and have different orders for sentence structure, which translation services used to overlook. But, they've come a long way.

Digital phone calls:

We all hear "this call may be recorded for training purposes," but rarely do we wonder what that entails. Turns out, these recordings may be used for training purposes, if a customer is aggrieved, but most of the time, they go into the database for an NLP system to learn from and improve in the future.

Data analysis:

Natural language capabilities are being integrated into data analysis workflows as more BI vendors offer a natural language interface to data visualizations. One example is smarter visual encodings, offering up the best visualization for the right task based on the semantics of the data.

Text analytics

Text analytics converts unstructured text data into meaningful data for analysis using different linguistic, statistical, and machine learning techniques.

Auto Correct and Auto prediction:

There are many softwares available nowadays that check grammar and spelling of the text we type and save us from embarrassing spelling and grammatical mistakes in our emails, texts or other documents.

Translation:


Social Media has brought the entire world together but with unity comes challenges like language barrier. With different translating softwares that work individually or are integrated within other applications, this hurdle has been easily defeated.

[2] Coherence resolution in detail.

Coreference resolution

Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.

Example:



"I", "my", and "she" belong to the same cluster and "Obama" and "he" belong to the same cluster.

- Coreference resolution resolution is the task of determining whether two mentions corefer, by which we mean they refer to the same entity in the discourse model (the same discourse entity).
- The set of corefering expressions is often called a coreference chain or a cluster

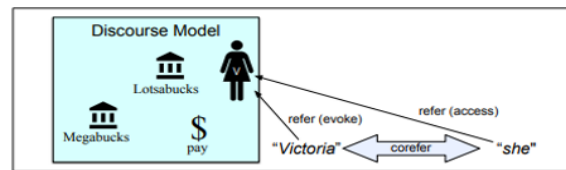


Figure 22.1 How mentions evoke and access discourse entities in a discourse model.

- For example, in abv figure processing ,
 - a coreference resolution algorithm would need to **find** at least four coreference chains,
 - corresponding to the four entities in the discourse model in Fig.
 - 1. {Victoria Chen, her, the 38-year-old, She}
 - 2. {Megabucks Banking, the company, Megabucks}
 - 3. {her pay}
 - 4. {Lotsabucks}
- Coreference resolution thus comprises two tasks : (1) identifying the mentions, and (2) clustering them into coreference chains/discourse entities

Coreference Phenomena: Linguistic Background

- Types of Referring Expressions

Indefinite Noun Phrases: The most common form of indefinite reference in English is marked with the determiner *a* (or *an*), but it can also be marked by a quantifier such as *some* or even the determiner *this*. Indefinite reference generally introduces into the discourse context entities that are new to the hearer.

- (21.6) a. Mrs. Martin was so very kind as to send Mrs. Goddard *a beautiful goose*.
 b. He had gone round one day to bring her *some walnuts*.
 c. I saw *this beautiful cauliflower* today.

Definite Noun Phrases: Definite reference, such as via NPs that use the English article *the*, refers to an entity that is identifiable to the hearer. An entity can be identifiable to the hearer because it has been mentioned previously in the text and thus is already represented in the discourse model:

(21.7) It concerns a white stallion which I have sold to an officer. But the pedigree of *the white stallion* was not fully established.

Alternatively, an entity can be identifiable because it is contained in the hearer's set of beliefs about the world, or the uniqueness of the object is implied by the description itself, in which case it evokes a representation of the referent into the discourse model, as in (21.9):

(21.8) I read about it in the *New York Times*.

(21.9) Have you seen the car keys?

The aim is to recognize the great prestige that the marathon has and jump in with this great race."

Demonstrative Pronouns: Demonstrative pronouns *this* and *that* can appear either alone or as determiners, for instance, *this ingredient*, *that spice*:

(21.14) I just bought a copy of Thoreau's *Walden*. I had bought one five years ago. *That one* had been very tattered; *this one* was in much better condition.

Note that *this NP* is ambiguous; in colloquial spoken English, it can be indefinite, as in (21.6), or definite, as in (21.14).

Zero Anaphora: Instead of using a pronoun, in some languages (including Chinese, Japanese, and Italian) it is possible to have an anaphor that has no lexical realization at all, called a **zero anaphor** or zero pronoun, as in the following Italian and Japanese examples from Poesio et al. (2016):

(21.15) EN [John]_i went to visit some friends. On the way [he]_i bought some wine.
IT [Giovanni]_i andò a far visita a degli amici. Per via ϕ_i comprò del vino.
JA [John]_i-wa yujin-o houmon-sita. Tochu-de ϕ_i wain-o ka-tta.

or this Chinese example:

(21.16) [我] 前一会精神上太紧张。[0] 现在比较平静了
[I] was too nervous a while ago. ... [0] am now calmer.

Zero anaphors complicate the task of mention detection in these languages.

Names: Names (such as of people, locations, or organizations) can be used to refer

Linguistic Properties of the Coreference Relation

- **Number Agreement:**
- Referring expressions and their referents must generally agree in number;
- English *she/her/he/him/his/it* are singular, *we/us/they/them* are plural, and *you* is unspecified for number.
- So a plural antecedent like *the chefs* cannot generally corefer with a singular anaphor like *she*.
- However, algorithms cannot enforce number agreement too strictly.
- First, semantically plural entities can be referred to by either *it* or *they*:

Person Agreement: English distinguishes between first, second, and third person, and a pronoun's antecedent must agree with the pronoun in person. Thus a third person pronoun (*he, she, they, him, her, them, his, her, their*) must have a third person antecedent (one of the above or any other noun phrase). However, phenomena like quotation can cause exceptions; in this example *I, my*, and *she* are coreferent:

(21.32) "I voted for Nader because he was most aligned with my values," she said.

Gender or Noun Class Agreement: In many languages, all nouns have grammatical gender or noun class⁶ and pronouns generally agree with the grammatical gender of their antecedent. In English this occurs only with third-person singular pronouns, which distinguish between *male* (*he, him, his*), *female* (*she, her*), and *nonpersonal* (*it*) grammatical genders. Non-binary pronouns like *ze* or *hir* may also occur in more recent texts. Knowing which gender to associate with a name in text can be complex, and may require world knowledge about the individual. Some examples:

(21.34) Maryam has a theorem. It is exciting. (it=the theorem, not Maryam)

Binding Theory Constraints: The **binding theory** is a name for syntactic constraints on the relations between a mention and an antecedent in the same sentence (Chomsky, 1981). Oversimplifying a bit, **reflexive** pronouns like *himself* and *herself* corefer with the subject of the most immediate clause that contains them (21.35), whereas nonreflexives cannot corefer with this subject (21.36).

(21.35) Janet bought herself a bottle of fish sauce. [herself=Janet]

(21.36) Janet bought her a bottle of fish sauce. [her≠Janet]

Recency: Entities introduced in recent utterances tend to be more salient than those introduced from utterances further back. Thus, in (21.37), the pronoun *it* is more likely to refer to Jim's map than the doctor's map.

(21.37) The doctor found an old map in the captain's chest. Jim found an even older map hidden on the shelf. It described an island.

Grammatical Role: Entities mentioned in subject position are more salient than those in object position, which are in turn more salient than those mentioned in oblique positions. Thus although the first sentence in (21.38) and (21.39) expresses roughly the same propositional content, the preferred referent for the pronoun *he* varies with the subject—John in (21.38) and Bill in (21.39).

21.2 COREFERENCE TASKS AND DATASETS 729

Verb Semantics: Some verbs semantically emphasize one of their arguments, biasing the interpretation of subsequent pronouns. Compare (21.40) and (21.41).

(21.40) John telephoned Bill. He lost the laptop.

(21.41) John criticized Bill. He lost the laptop.

These examples differ only in the verb used in the first sentence, yet “he” in (21.40) is typically resolved to John, whereas “he” in (21.41) is resolved to Bill. This may be due to the link between implicit causality and saliency: the implicit cause of a “criticizing” event is its object, whereas the implicit cause of a “telephoning” event is its subject. In such verbs, the entity which is the implicit cause is more salient.

Selectional Restrictions: Many other kinds of semantic knowledge can play a role in referent preference. For example, the selectional restrictions that a verb places on its arguments (Chapter 10) can help eliminate referents, as in (21.42).

[3] Explain in detail about Sentiment Classification.

Classification lies at the heart of both human and machine intelligence. Deciding what letter, word, or image has been presented to our senses, recognizing faces or voices, sorting mail, assigning grades to homeworks; these are all examples of assigning a category to an input. The potential challenges of this task are highlighted by the fabulist Jorge Luis Borges (1964), who imagined classifying animals into:

(a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.

Many language processing tasks involve classification, although luckily our classes are much easier to define than those of Borges. In this chapter we introduce the naive Bayes algorithm and apply it to **text categorization**, the task of assigning a label or category to an entire text or document.

We focus on one common text categorization task, **sentiment analysis**, the extraction of **sentiment**, the positive or negative orientation that a writer expresses toward some object. A review of a movie, book, or product on the web expresses the author's sentiment toward the product, while an editorial or political text expresses sentiment toward a candidate or political action. Extracting consumer or public sentiment is thus relevant for fields from marketing to politics.

The simplest version of sentiment analysis is a binary classification task, and the words of the review provide excellent cues. Consider, for example, the following phrases extracted from positive and negative reviews of movies and restaurants. Words like *great*, *richly*, *awesome*, and *pathetic*, and *awful* and *ridiculously* are very informative cues:

- + ...zany characters and richly applied satire, and some great plot twists
- It was pathetic. The worst part about it was the boxing scenes...
- + ...awesome caramel sauce and sweet toasty almonds. I love this place!
- ...awful pizza and ridiculously overpriced...

spam detection

Spam detection is another important commercial application, the binary classification task of assigning an email to one of the two classes *spam* or *not-spam*. Many lexical and other features can be used to perform this classification. For example you might quite reasonably be suspicious of an email containing phrases like “online pharmaceutical” or “WITHOUT ANY COST” or “Dear Winner”.

language id

Another thing we might want to know about a text is the language it's written in. Texts on social media, for example, can be in any number of languages and we'll need to apply different processing. The task of **language id** is thus the first step in most language processing pipelines. Related text classification tasks like **authorship attribution**—determining a text's author—are also relevant to the digital humanities, social sciences, and forensic linguistics.

authorship attribution

4.4 Optimizing for Sentiment Analysis

While standard naive Bayes text classification can work well for sentiment analysis, some small changes are generally employed that improve performance.

First, for **sentiment classification** and a number of other text classification tasks, whether a word occurs or not seems to matter more than its frequency. Thus it often improves performance to clip the word counts in each document at 1 (see the end of the chapter for pointers to these results). This variant is called **binary**

multinomial naive Bayes or **binary NB**. The variant uses the same Eq. 4.10 except that for each document we remove all duplicate words before concatenating them into the single big document. Fig. 4.3 shows an example in which a set of four documents (shortened and text-normalized for this example) are remapped to binary, with the modified counts shown in the table on the right. The example is worked without add-1 smoothing to make the differences clearer. Note that the results counts need not be 1; the word *great* has a count of 2 even for Binary NB, because it appears in multiple documents.

Four original documents:		NB Counts		Binary Counts	
		+	-	+	-
- it was pathetic the worst part was the boxing scenes	and	2	0	1	0
- no plot twists or great scenes	boxing	0	1	0	1
+ and satire and great plot twists	film	1	0	1	0
+ great scenes great film	great	3	1	2	1
	it	0	1	0	1
	no	0	1	0	1
	or	0	1	0	1
	part	0	1	0	1
	pathetic	0	1	0	1
	plot	1	1	1	1
	satire	1	0	1	0
	scenes	1	2	1	2
	the	0	2	0	1
	twists	1	1	1	1
	was	0	2	0	1
	worst	0	1	0	1
After per-document binarization:					
- it was pathetic the worst part boxing scenes					
- no plot twists or great scenes					
+ and satire great plot twists					
+ great scenes film					

Figure 4.3 An example of binarization for the binary naive Bayes algorithm.

A second important addition commonly made when doing text classification for sentiment is to deal with negation. Consider the difference between *I really like this movie* (positive) and *I didn't like this movie* (negative). The negation expressed by *didn't* completely alters the inferences we draw from the predicate *like*. Similarly, negation can modify a negative word to produce a positive review (*don't dismiss this film, doesn't let us get bored*).

A very simple baseline that is commonly used in sentiment analysis to deal with negation is the following: during text normalization, prepend the prefix *NOT_* to every word after a token of logical negation (*n't, not, no, never*) until the next punctuation mark. Thus the phrase

didn't like this movie , but I

becomes

didn't NOT_like NOT_this NOT_movie , but I

Newly formed 'words' like *NOT_like*, *NOT_recommend* will thus occur more often in negative document and act as cues for negative sentiment, while words like *NOT_bored*, *NOT_dismiss* will acquire positive associations. We will return in Chapter 16 to the use of parsing to deal more accurately with the scope relationship between these negation words and the predicates they modify, but this simple baseline works quite well in practice.

Finally, in some situations we might have insufficient labeled training data to train accurate naive Bayes classifiers using all words in the training set to estimate positive and negative sentiment. In such cases we can instead derive the positive

sentiment
lexicons
General
Inquirer
LIWC

and negative word features from **sentiment lexicons**, lists of words that are pre-annotated with positive or negative sentiment. Four popular lexicons are the **General Inquirer** (Stone et al., 1966), **LIWC** (Pennebaker et al., 2007), the opinion lexicon of Hu and Liu (2004a) and the MPQA Subjectivity Lexicon (Wilson et al., 2005).

For example the MPQA subjectivity lexicon has 6885 words, 2718 positive and 4912 negative, each marked for whether it is strongly or weakly biased. Some samples of positive and negative words from the MPQA lexicon include:

+ : *admirable, beautiful, confident, dazzling, ecstatic, favor, glee, great*
- : *awful, bad, bias, catastrophe, cheat, deny, envious, foul, harsh, hate*

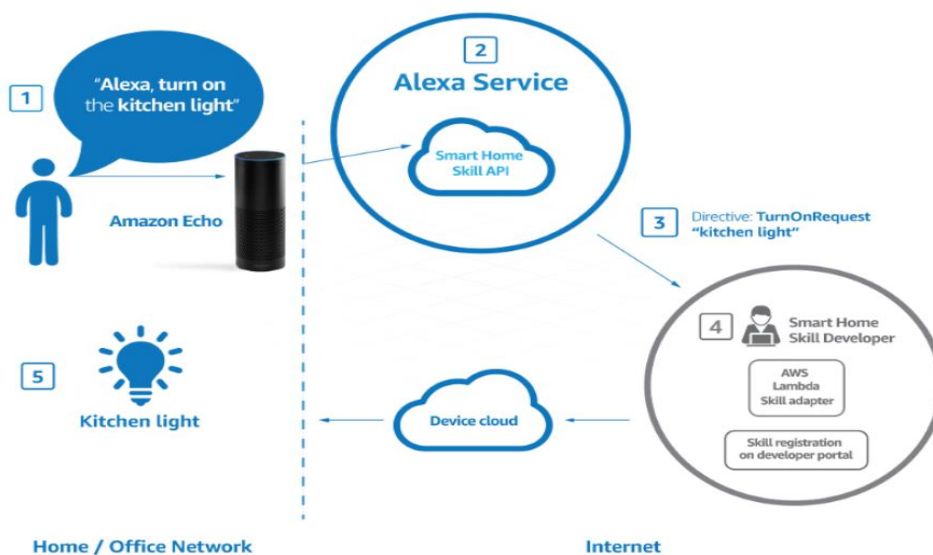
A common way to use lexicons in a naive Bayes classifier is to add a feature that is counted whenever a word from that lexicon occurs. Thus we might add a feature called 'this word occurs in the positive lexicon', and treat all instances of words in the lexicon as counts for that one feature, instead of counting each word separately. Similarly, we might add as a second feature 'this word occurs in the negative lexicon' of words in the negative lexicon. If we have lots of training data, and if the test data matches the training data, using just two features won't work as well as using all the words. But when training data is sparse or not representative of the test set, using dense lexicon features instead of sparse individual-word features may generalize better.

We'll return to this use of lexicons in Chapter 20, showing how these lexicons can be learned automatically, and how they can be applied to many other tasks beyond **sentiment classification**.

[4] Is alexa example of nlp explain ,mention some.

Ans: According to Adi Agashe, Program Manager at Microsoft, Alexa is built based on natural language processing (NLP), a procedure of converting speech into words, sounds, and ideas. Amazon records your words.

Amazon records your words. Indeed, interpreting sounds takes up a lot of computational power, the recording of your speech is sent to Amazon's servers to be analyzed more efficiently.



Source

Alexa is also a part of Amazon which uses Machine Learning to predict what the user will ask the information for and then based on that gives a rich user experience when the user asks the question.

More examples:

- Google Natural Language API
- The Google Natural Language API is an easy to use interface to a set of powerful NLP models which have been pre-trained by Google to perform various tasks.
- As these models have been trained on enormously large document corpuses, their performance is usually quite good as long as they are used on datasets that do not make use of a very idiosyncratic language.
- Voice-enabled applications such as Alexa, Siri, and Google Assistant use NLP and Machine Learning (ML) to answer our questions, add activities to our calendars and call the contacts that we state in our voice commands.
- NLP is not only making our lives easier, but revolutionizing the way we work, live, and play.

[5] How does nlp advantageous in day to day life, name of the applications which you see in daily life, is nlp part of AI justify.

Ans:

- NLP makes sense of the human languages by using machine learning. Its ultimate objective is to process written and spoken information to derive meaning to improve our personal and professional lives.
- Among the many advantages of NLP, is the ability to present unstructured data into comprehensive information, like enhancing customer experiences facilitated by virtual assistants that talk, think and act like humans.
- By using NLP, you can better extract data or information from text-based documents and improve on more complex analytics tasks like sentiment analysis.

Let's take a look at the most interesting applications of natural language processing in business:

- Sentiment Analysis.
- Text Classification.
- Chatbots & Virtual Assistants.
- Text Extraction.
- Machine Translation.
- Text Summarization.
- Market Intelligence.
- Auto-Correct.

“NLP makes it possible for humans to talk to machines:” This branch of AI enables computers to understand, interpret, and manipulate human language. Like machine learning or deep learning, NLP is a subset of AI.

Natural language processing (NLP) is a branch of artificial intelligence within computer science that focuses on helping computers to understand the way that humans write and speak.

This is a difficult task because it involves a lot of unstructured data. The style in which people talk and write (sometimes referred to as ‘tone of voice’) is unique to individuals, and constantly evolving to reflect popular usage.

Understanding context is also an issue – something that requires semantic analysis for machine learning to get a handle on it.

[6] Advantages and Disadvantages of nlp ,challenges faced in nlp,write short note on machine translation.

Advantages of NLP

- NLP helps users to ask questions about any subject and get a direct response within seconds.
- NLP offers exact answers to the question.
- It does not offer unnecessary and unwanted information.
- NLP helps computers to communicate to humans(Any languages).

- It is very time efficient.
- NLP improves
- The efficiency of documentation processes.
- Accuracy of documentation.
- Identify the information from large databases.

Disadvantages of NLP

- May not show context.
- Unpredictable.
- This require more keystrokes.
- NLP is unable to adapt to the new domain.
- NLP has a limited function.
- NLP is built for a single and specific task.

Here are the major challenges around NLP that one must be aware of.

- Training Data. NLP is mainly about studying the language and to be proficient, it is essential to spend a substantial amount of time listening, reading, and understanding it. ...
- Development Time. ...
- Homonyms. ...
- Misspellings. ...
- False Positives.

Machine translation (MT) is automated, meaning it's the translation of text by a computer with no human involvement.

It works by using computer software to translate text from one language (source language) to another language (target language).

Machine Translation (MT) is the task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language.

While machine translation is one of the oldest subfields of artificial intelligence research, the recent shift towards large-scale empirical techniques has led to very significant improvements in translation

quality. The Stanford Machine Translation group's research interests lie in techniques that utilize both statistical methods and deep linguistic analyses.

Research in our group currently focuses on the following topics:

- Better Training in MT
- Chinese MT
- Arabic MT