Unit 5. Introduction to natural language processing

Estimated time

00:30

Overview

This unit introduces Natural Language Processing. It covers key applications of NLP, basics concepts and terminology, tools and services and NLP challenges.

IBM Training

Unit objectives

- Explain what natural language processing is.
- Identify NLP use cases.
- Explain basic NLP concepts and terminology.
- List the tools and services for NLP.

Figure 5-1. Unit objectives

5.1. Natural language processing overview

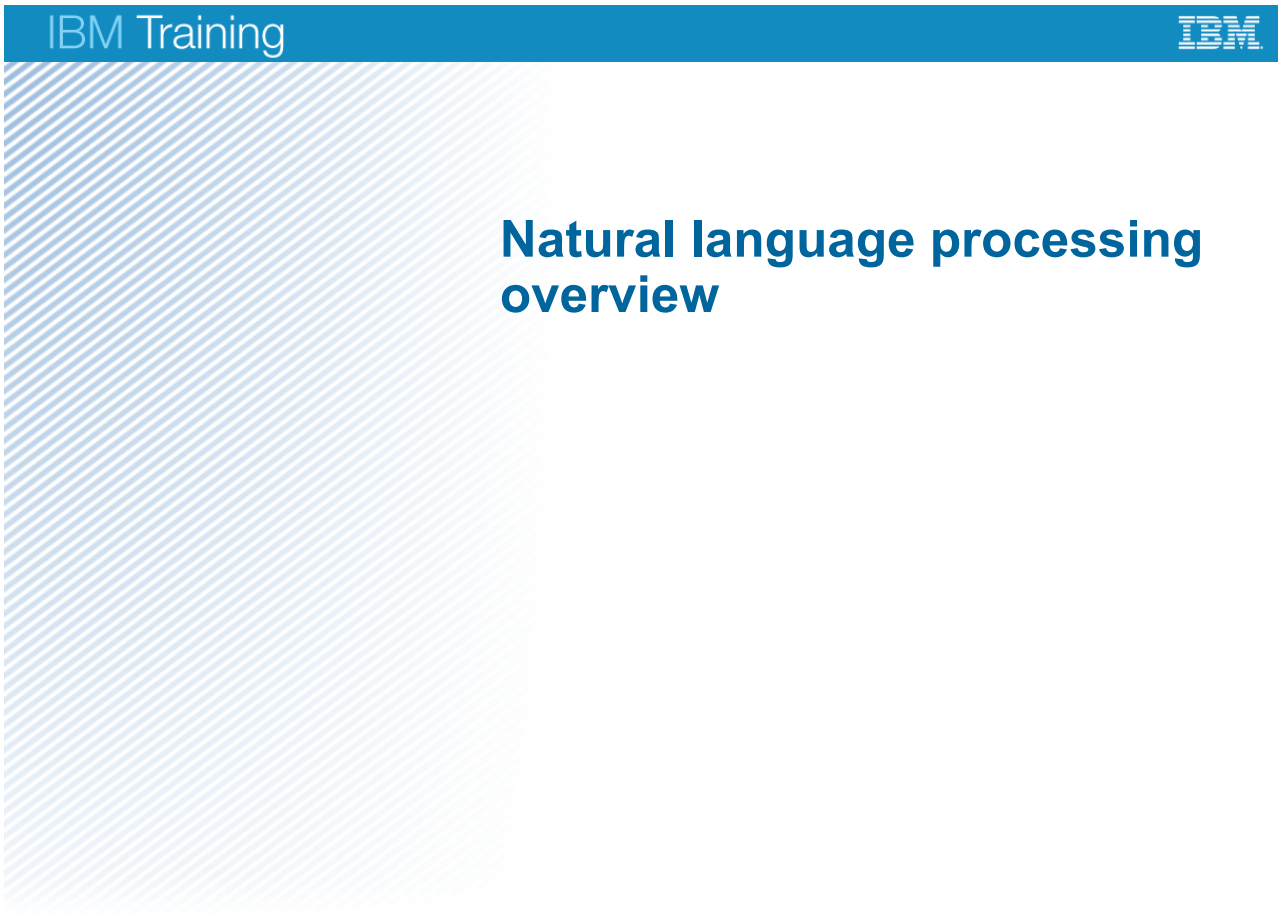


Figure 5-2. Natural language processing overview

IBM Training

Topics

- Natural language processing overview
- Natural language processing use cases
- Natural language processing basic concepts and terminology
- Natural language processing tools and services

Figure 5-3. Topics

IBM Training

What is natural language processing

- NLP is the study of the computational treatment of natural (human) language.
- It enables machines to understand human communication to extract different information.
- Examples of NLP applications: Analysis of text in emails, human speech, social media, or optical character recognition (OCR) from documents (text that is scanned from actual documents).
- NLP has its origins in machine translation from the 1950s.
- NLP advanced over the years by combining the power of artificial intelligence (AI), computational linguistics, and computer science.

Figure 5-4. What is natural language processing

NLP stands for natural language processing. It is a subfield of computer science and AI concerned with the processing human language by computers. It is one of the most important fields in computer science in both industry and academia.

NLP enables machines to understand human communication to extract different information.

Examples of NLP applications include analysis of text in emails, human speech, social media, or optical character recognition (OCR) from documents (text that is scanned from actual documents).

NLP has its origins in machine translation from the 1950s.

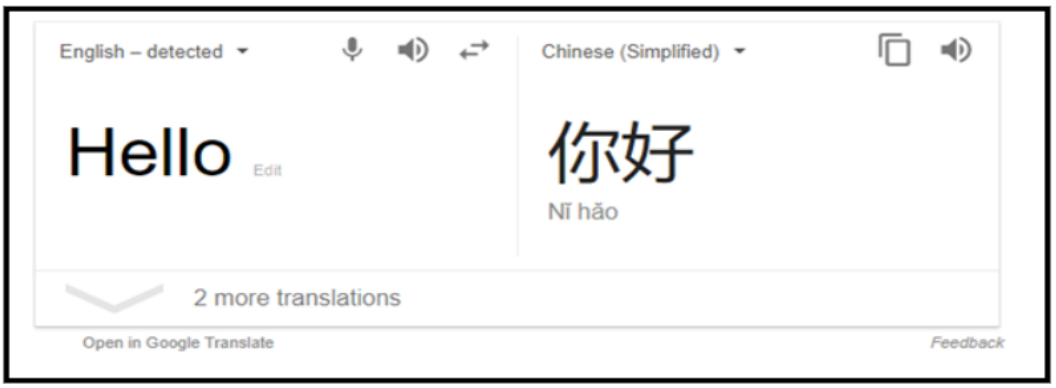
The first machine translation was from English to Russian and vice versa, but with poor and inaccurate results.

Machine translation and other NLP applications advanced over the years by combining the power of artificial intelligence (AI), computational linguistics, and computer science.

IBM Training

Natural language processing applications

- Machine translation
- Information retrieval: Search engines
- Spell checkers
- Natural language assistants



Google Translate

Figure 5-5. Natural language processing applications

Here are the most popular NLP tasks:

- Machine translation: Automatically translating one language to another
- Information retrieval: Search engines, such as Google and Bing
- Spell checkers
- Natural language assistants, such as Siri and Alexa

Natural language processing challenges

- **Domains:** Higher accuracy for specific domains compared to generic domains.
- **Language:** English gets the most attention because it is an international language.
- **Medium:** Processing speech is more difficult than processing text.

You can understand your NLP problem by focusing on the following areas:

- Become familiar with your data.
- Understand the challenges of your particular use case.
- Review the state-of-the-art solutions and technologies for similar problems.

Figure 5-6. Natural language processing challenges

There are always some challenges that need to be tackled for any case. In NLP, here are the most popular challenges:

- **Domains**: Higher accuracy for specific domains compared to generic domains.
- **Language:** English gets the most attention because it is an international language.
- **Medium:** Processing speech is more difficult than processing text.

To resolve some of these challenges, you must become familiar with your data and understand the challenges of your particular use case. Think about how you will acquire the data and how to validate its quality. Think about the deployment of your solution and how you are planning to cover all these points.

Finally, review the state-of-the-art solutions and technologies for similar problems and how these issues were resolved.

5.2. Natural language processing use cases

IBM Training

Natural language processing use cases

Figure 5-7. Natural language processing use cases

IBM Training

Topics

- Natural language processing overview
- Natural language processing use cases
- Natural language processing basic concepts and terminology
- Natural language processing tools and services

Figure 5-8. Topics

IBM Training

Natural language processing use cases

Information extraction

- **Goal:** Parse the input text to extract valuable output.
  - Examples: Entity extraction, relation identification extraction, text summarization
- **Unstructured text**: Dynamic structure (for example emails, newspaper articles, and user reviews).
- **Structured text:** Defined structure (for example, a database table).

Figure 5-9. Natural language processing use cases

NLP technology can be used to extracting information from unstructured text such as emails, newspaper articles, and user reviews into structured text.

Entity extraction refers to extracting *entities* from the text such as organizations, people, locations and so on. For example, the World Health Organization, IBM, Sara, John, Paris, US.

Relation extraction refers to identifying the relationship between entities, for example, “Abraham Lincoln was a US president”; “Ginni Rometty is the CEO of IBM”.

Text summarization refers to the technique of shortening long pieces of text. Automatic text summarization is a common use case in machine learning and natural language processing.

Structured text mostly takes the form of tables or values in a structured form.

The goal of information extraction is to parse the incoming text, identify important mentions and their relations, and extract valuable output into structured text. Doing so can be used to automate the process of reading articles and passages to convert this information into a structured format. Computer systems can then manage this information and take proper actions.

Natural language processing use cases (cont.)

Sentiment analysis

- The process of identifying emotions or opinions that are expressed in user input.
- This analysis is used in marketing, retention plans, and emotional intelligence for chatbots.

Figure 5-10. Natural language processing use cases (cont.)

Sentiment analysis is the process of identifying emotions or opinions that are expressed in user input.

Sentiment analysis answers various questions, such as how people feel about your product or whether your customers are satisfied with your customer service.

It is used in marketing and retention plans, and emotional intelligence for chatbots, that is, it enables chatbots to direct the conversation.

Machine learning algorithms brought many advances to this field and are still improving.

Natural language processing use cases (cont.)

- **Question and answering**: Building solutions that can answer questions that are asked by humans in natural language.
  - Examples: Natural language questions used to retrieve answers from forums; FAQ application; chatbots
- **Speech recognition:** Converts spoken language into text.
  - Example: Chatbot interactive talk

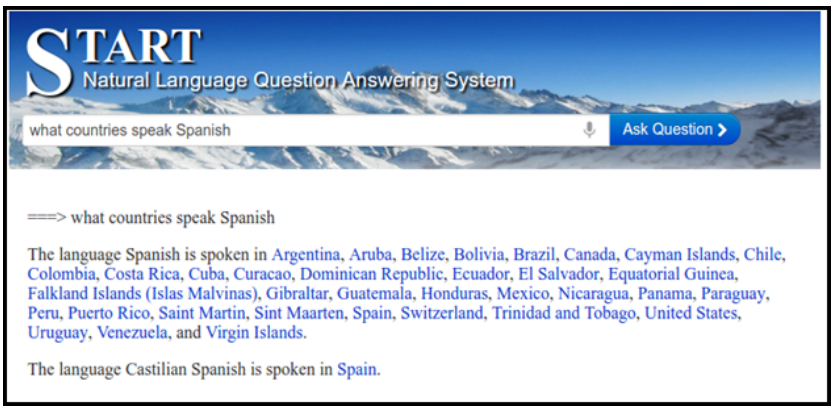


Figure 5-11. Natural language processing use cases (cont.)

Question and answering

Building solutions that can answer questions that are asked by humans in natural language. A question and answering system can be used for the following tasks:

- Retrieving answers from forums.
- Building a Frequently Asked Questions (FAQs) system.
- Training chatbots

Speech recognition is another use case that helps advancing the capabilities of many different applications. It converts spoken language into text. It can be used in many applications in several domains, such as having an interactive talk with a chatbot. It can also be used in Internet of Things (IoT) applications.

5.3. Natural language processing basic concepts and terminology

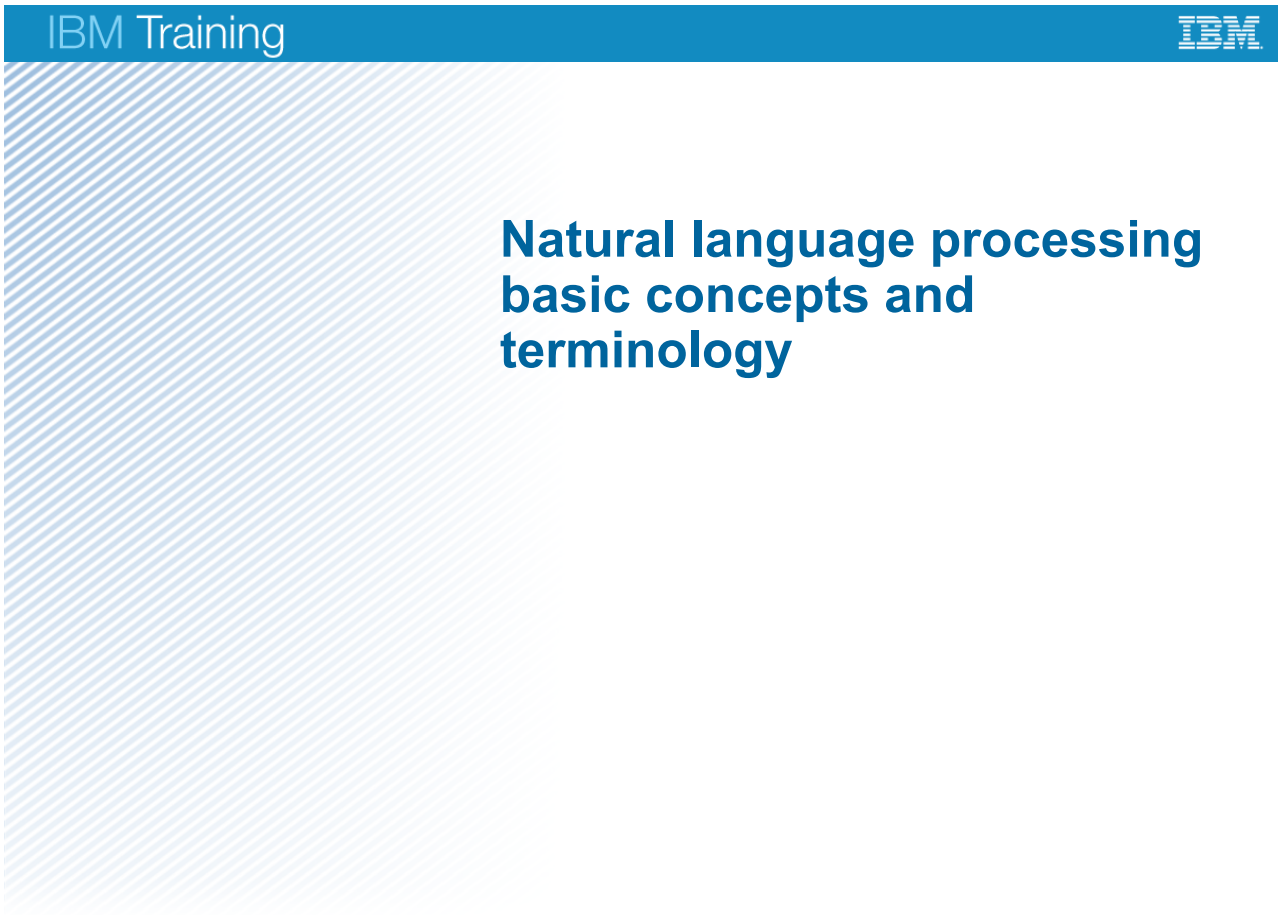


Figure 5-12. Natural language processing basic concepts and terminology

This section introduces basic concepts and terminologies such as synonymy, polysemy, hyponymy, and hypernymy. The taxonomy for similar concepts has applications in the education and machine learning fields because they rely on word-sense disambiguation.

IBM Training

Topics

- Natural language processing overview
- Natural language processing use cases
- Natural language processing basic concepts and terminology
- Natural language processing tools and services

Figure 5-13. Topics

IBM Training

Basic concepts and terminologies

- **Synonyms:** Words that are written differently but are similar in meaning.
  - Example: Clever and smart
- **Antonyms:** Words that have meanings that are opposite to each other.
  - Example: Clever and stupid
- **Usage example:** In information retrieval, you might want to expand the keywords search by retrieving the synonyms of the query words.

Figure 5-14. Basic concepts and terminologies

Synonyms are words that are written differently but are similar in meaning. For example:

- Clever and smart
- Begin and start
- Beautiful and pretty
- Sad and unhappy

Antonyms are words that have meanings that are opposite to each other. For example:

- Clever and stupid
- Begin and end
- Beautiful and ugly
- Sad and happy

Usage example: In information retrieval, you might want to expand the keywords search by retrieving the synonyms of the query words.

Basic concepts and terminologies (cont.)

Homonyms: Words that have the same written form but have unrelated meanings. There are two types of homonyms:

- **Homographs:** Words that have the same written form. For example:
  - This answer is **right**.
  - The building is on the **right** side of the river.
  - You have the **right** to remain silent.
  - Come here **right** now.
- **Homophones:** Words that sound similar when spoken but have different meanings and spellings. For example:
  - “left” and “lift”.
  - “right” and “write”.

Figure 5-15. Basic concepts and terminologies (cont.)

Homonyms are words that have the same written form but have unrelated meanings. There are two types of homonyms:

- Homographs
- Homophones

Homographs are words that have the same written form. For example:

- This answer is ***right***.
- The building is on the ***right*** side of the river.
- You have the ***right*** to remain silent.
- Come here ***right*** now.

Although the word ***right*** has the same written form in the examples, you notice the difference between the meanings in each sentence.

Homophones are words that sound similar when spoken but have different meanings and spellings. For example:

- “left” and “lift”.
- “right” and “write”.

Basic concepts and terminologies (cont.)

- Homonyms **challenges**:
  - How do you translate “right” so that it has the correct meaning?
  - How do you differentiate two words that sound similar when you convert speech to text?

Figure 5-16. Basic concepts and terminologies (cont.)

Homonyms introduce challenges into NLP operations such as machine translation and speech recognition. How do you translate *right* so that it has the correct meaning? How do you differentiate two words that sound similar when you convert speech to text?

Basic concepts and terminologies (cont.)

- **Polysemy:** Words that have the same written form and a related meaning. For example:
  - You must face your fear.
  - Her face is beautiful.
- **Hyponymy:** A word is a hyponym of another word if it represents a subclass of the other word. For example:
  - Orange is a hyponym of fruit.
  - Yellow is a hyponym of color.

Figure 5-17. Basic concepts and terminologies (cont.)

Polysemy refers to words that have the same written form and related meaning. For example:

- You must *face* your fear.
- Her face is beautiful.

Hyponymy: A word is a hyponym of another word if it represents a subclass of the other word. For example:

- Orange is a hyponym of fruit.
- Yellow is a hyponym of color.

Basic concepts and terminologies (cont.)

- **Hypernymy:** One word is the hypernym of another word if it represents a superclass of the other word. For example:
  - Fruit is a hypernym of orange.
  - Color is a hypernym of yellow.
- **Usage example:** Comparing the semantic similarity.

Figure 5-18. Basic concepts and terminologies (cont.)

Hypernymy: One word is the hypernym of another word if it represents a superclass of the other word. For example:

- *Fruit* is a hypernym of orange.
- *Color* is a hypernym of yellow.

Usage example: Comparing the semantic similarity.

5.4. Natural language processing tools and services

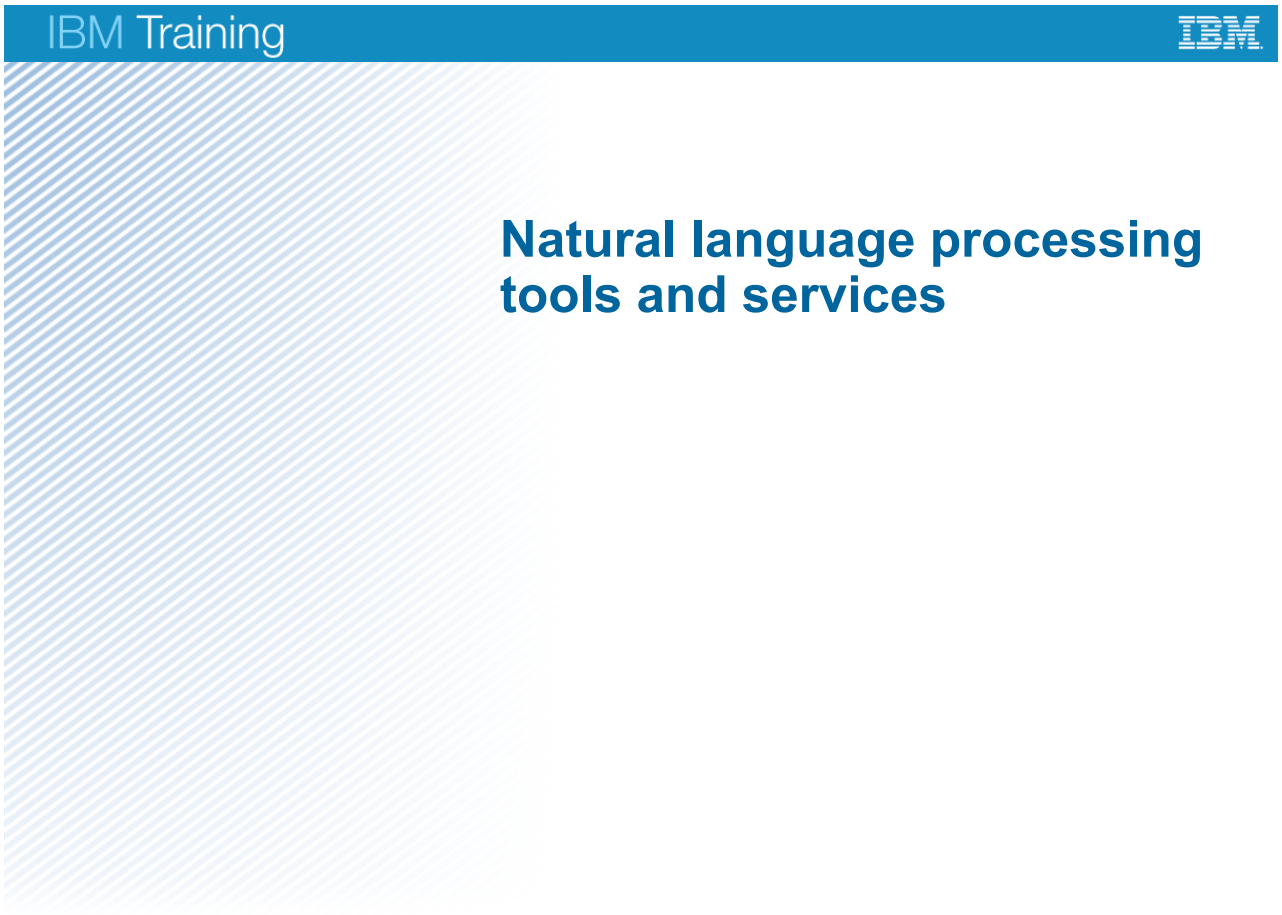


Figure 5-19. Natural language processing tools and services

IBM Training

Topics

- Natural language processing overview
- Natural language processing use cases
- Natural language processing basic concepts and terminology
- Natural language processing tools and services

Figure 5-20. Topics

IBM Training

Natural language processing tools and services

Open-source NLP tools:

- **Apache OpenNLP:** Provides tokenizers, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, co-reference resolution, and more.
- **Stanford Core NLP:** A suite of NLP tools that provide part-of-speech tagging, a named entity recognizer, a co-reference resolution system, sentiment analysis, and more.
- **Natural Language Toolkit (NLTK):** A Python library that provides modules for processing text, classifying, tokenizing, stemming, tagging, parsing, and more.
- **WordNet:** One of the most popular lexical databases for the English language. Supported by various API and programming languages.

Figure 5-21. Natural language processing tools and services

There are many open source tools that you can use for NLP. For example:

- Open NLP that is based on Java. It provides many functions for text processing, such as tokenizers, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, co-reference resolution, and more. For more information, see https://opennlp.apache.org/
- Stanford Core NLP, which is written in Java. It is a suite of NLP tools that provide part-of-speech tagging, a named entity recognizer, a co-reference resolution system, sentiment analysis, and more. It supports many languages, such as English, German, French, Arabic, Spanish, and Chinese. For more information, see https://stanfordnlp.github.io/CoreNLP/.
- NLTK provides the same processes as the other NLP suites, but in the Python language. For more information, see https://www.nltk.org/.
- WordNet is a popular lexical database that is used in research. There are many APIs and languages that you can use to access WordNet. For example, you can make a call to retrieve a synonym of a word. WordNet is available online and as an offline version that you can download. For more information, see https://wordnet.princeton.edu/.

There are other libraries, such as Unstructured Information Management Architecture (UIMA). IBM Watson uses UIMA to analyze unstructured data. The Apache Clinical Text Analysis and Knowledge Extraction System (Apache cTAKES) is a UIMA-based system that is used to extract information from medical records.

IBM Training

Natural language processing tools and services (cont.)

Services:

- Examples: IBM Cloud, Microsoft Cloud (Azure), and Google Cloud
- IBM offers its AI services through IBM Cloud. The NLP services that are provided include the following ones (among others):
  - Watson Natural Language Classifier for text classification
  - Watson Natural Language Understanding for entity identification and relation extraction

Figure 5-22. Natural language processing tools and services (cont.)

Instead of using low-level libraries, you can use many cloud services that accomplish high-level NLP tasks, for example, IBM Cloud, Microsoft Cloud (Azure), and Google Cloud.

IBM offers its AI services through IBM Cloud. The NLP services that are provided include the following ones (among others):

- Watson Natural Language Classifier for text classification
- Watson Natural Language Understanding for entity identification and relation extraction

You can rapidly build a complete NLP application by using these services.

IBM Training

Unit summary

- Explain what natural language processing is.
- Identify NLP use cases.
- Explain basic NLP concepts and terminology.
- List the tools and services for NLP.

Figure 5-23. Unit summary