

UNIT 2

Knowledge Acquisition

(8 Hours) — Knowledge Engineering (CACS458)

⚠ Please Note

These notes are generated to help with revision and **may contain errors or omissions**. Please verify important definitions, formulas and examples against your textbook, lecture notes and the official TU syllabus before relying on them in an exam.

2.1 Information Gathering

Knowledge acquisition is the process of extracting, structuring and organising knowledge from sources (human experts, documents, databases, the web) so it can be placed in a knowledge base. **Information gathering** is the first step: collecting the raw material from which knowledge is built.

Sources of Knowledge

- Domain experts, practitioners and users.
- Books, manuals, research papers, reports and standards.
- Databases, logs and data warehouses.
- The web, social media and public datasets.
- Sensors and observations of real work.

Techniques for Gathering Knowledge from Experts

Technique	Description	Strength / Weakness
Interview	Structured, unstructured or semi-structured questioning of the expert	Rich detail / time-consuming, expert may omit tacit steps
Observation	Watching the expert perform tasks in the real setting	Captures actual practice / experts may behave differently when watched
Protocol analysis (think-aloud)	Expert speaks thoughts while solving a problem	Reveals reasoning / disturbs normal thinking
Questionnaire / survey	Written questions to many people	Cheap and wide / shallow answers

Technique	Description	Strength / Weakness
Card sorting & repertory grid	Expert groups concepts or compares items	Reveals concept structure / needs training
Document analysis	Reading manuals, cases, regulations	Quick start / may be outdated
Machine learning / data mining	Learning rules automatically from data	Scalable / needs good quality data

Note: The knowledge acquisition bottleneck arises because experts often cannot easily explain their own tacit knowledge, and the process is slow and costly.

Challenges in Information Gathering

- Experts are busy, or reluctant to share knowledge.
- Knowledge is tacit, incomplete, inconsistent or uncertain.
- Vocabulary differences between expert and knowledge engineer.
- Information overload and low-quality sources.

2.2 Information Retrieval

Information retrieval (IR) is the process of finding documents or pieces of information that are relevant to a user's *information need*, usually expressed as a query, from a large collection. Web search engines are the most common IR systems.

Basic IR Process

- **Crawling / collection** — gather documents.
- **Pre-processing** — tokenisation, lower-casing, stop-word removal, stemming.
- **Indexing** — build an inverted index mapping each term to the documents containing it.
- **Query processing** — apply the same pre-processing to the query.
- **Matching & ranking** — score documents against the query and sort them.
- **Evaluation** — measure quality using precision and recall.

Retrieval Models

Model	Idea
Boolean model	Query uses AND, OR, NOT; document either matches or does not. No ranking.

Model	Idea
Vector space model	Documents and query are vectors of term weights; ranking by cosine similarity.
Probabilistic model	Ranks documents by probability of being relevant to the query.
Link-based (PageRank)	Uses hyperlink structure to estimate page importance.

TF-IDF Weighting

In the vector space model, the weight of term t in document d is $w(t, d) = \mathbf{tf}(t, d) \times \mathbf{idf}(t)$, where $\mathbf{tf}$ is the term frequency in the document and $\mathbf{idf}(t) = \log(N / \mathbf{df}(t))$, N being the total number of documents and $\mathbf{df}(t)$ the number of documents containing t . Terms that are frequent in a document but rare in the collection get a high weight.

Cosine similarity between query q and document d is: $\cos(q, d) = (q \cdot d) / (|q| \times |d|)$.

Evaluation Measures

Precision = relevant retrieved $\div$ total retrieved. **Recall** = relevant retrieved $\div$ total relevant in collection. **F-measure** = $2PR / (P + R)$.

Example 2.1 — Precision and recall

A collection has 20 relevant documents for a query. The system retrieves 25 documents, of which 15 are relevant. Find precision, recall and F-measure.

Solution

$$\text{Precision} = 15 / 25 = 0.60$$

$$\text{Recall} = 15 / 20 = 0.75$$

$$\text{F-measure} = 2 \times 0.60 \times 0.75 / (0.60 + 0.75) = 0.90 / 1.35 \approx 0.667$$

Example 2.2 — TF-IDF

$N = 1000$ documents. The term “ontology” appears in 10 of them, and appears 3 times in document D1. Use log base 10 and compute the TF-IDF weight of “ontology” in D1 (raw tf).

Solution

$$\text{idf} = \log_{10}(1000 / 10) = \log_{10}(100) = 2$$

$$\text{tf-idf} = 3 \times 2 = 6$$

2.3 Applications of Natural Language Processing (NLP)

Natural language processing enables computers to understand, interpret and generate human language. In knowledge engineering, NLP is a major tool for acquiring knowledge from unstructured text.

Typical NLP Processing Pipeline



Figure 2.1: A typical NLP processing pipeline.

Applications

- Machine translation (e.g. English $\rightleftharpoons$ Nepali).
- Information extraction and knowledge-base population from text.
- Question answering and chatbots.
- Text summarisation and sentiment analysis.
- Speech recognition and spelling/grammar checking.
- Search engines and document classification.

2.3.1 Morphology, Lexicon, Syntax and Semantics

Level	Concerned with	Example
Morphology	Structure and formation of words (roots, prefixes, suffixes, inflection)	“unhappiness” = un + happy + ness; “books” = book + s
Lexicon	Dictionary of words with their categories and meanings	run — verb, noun; has several senses
Syntax	How words combine into grammatical sentences	“The cat sat on the mat” is valid; “Cat the mat on sat” is not
Semantics	Meaning of words and sentences	“Colourless green ideas sleep furiously” is syntactically fine but semantically odd
Pragmatics	Meaning in context and intention of speaker	“Can you pass the salt?” is a request, not a question about ability

Stemming chops a word to a crude stem (“studies” → “studi”), while **lemmatisation** returns the dictionary form using the lexicon (“studies” → “study”).

2.3.2 Parsing

Parsing is the process of analysing the grammatical structure of a sentence according to a grammar. A **context-free grammar (CFG)** consists of production rules such as $S \rightarrow NP VP$, $NP \rightarrow Det N$, $VP \rightarrow V NP$.

- **Constituency parsing** — builds a phrase-structure tree (NP, VP, ...).
- **Dependency parsing** — links words by head-dependent relations (subject, object).
- **Top-down parsing** starts from S and expands; **bottom-up parsing** starts from words and reduces; chart parsers (CYK, Earley) avoid repeated work.
- **Ambiguity** — one sentence can have more than one parse (“I saw the man with a telescope”).

Example 2.3 — Parse of “The dog chased a cat”

Grammar: $S \rightarrow NP VP$; $NP \rightarrow Det N$; $VP \rightarrow V NP$; $Det \rightarrow the \mid a$; $N \rightarrow dog \mid cat$; $V \rightarrow chased$.

Parse tree (bracketed)

[S [NP [Det The] [N dog]] [VP [V chased] [NP [Det a] [N cat]]]]

2.3.3 Part-of-Speech (POS) Tagging

POS tagging assigns each word a grammatical category such as noun (NN), verb (VB), adjective (JJ), determiner (DT), preposition (IN). The difficulty is that many words are ambiguous (“book” can be noun or verb). Methods include rule-based taggers, **Hidden Markov Model** taggers, and neural taggers.

Example 2.4 — POS tagging

Sentence: *Ram reads good books.*

Tags

Ram/NNP reads/VBZ good/JJ books/NNS ./.

2.3.4 Named Entity Recognition (NER) / Tagging

Named entity recognition locates and classifies named entities in text into categories such as PERSON, ORGANISATION, LOCATION, DATE and MONEY. It is essential for information extraction and for populating knowledge bases.

Example 2.5 — NER

“Sita joined Tribhuvan University in Kathmandu in 2021.”

Entities

Sita → PERSON; Tribhuvan University → ORGANISATION; Kathmandu → LOCATION; 2021 → DATE

Note: Common NER approaches: dictionary/gazetteer lookup, rule-based patterns, statistical models (CRF, HMM) and deep-learning models. IOB tagging marks Beginning, Inside and Outside of entities.

Key Takeaways — Unit 2

- ✓ Knowledge acquisition collects and structures knowledge from experts, documents and data; interviews, observation and protocol analysis are key techniques.
- ✓ Information retrieval finds relevant documents using indexing and ranking (Boolean, vector space, probabilistic); precision and recall measure quality.
- ✓ $TF-IDF = tf \times \log(N/df)$ gives high weight to terms frequent in a document but rare in the collection.
- ✓ NLP levels: morphology, lexicon, syntax, semantics (and pragmatics).
- ✓ Parsing gives sentence structure; POS tagging labels word categories; NER finds entities such as persons, places and organisations.

— End of Unit 2 —