Entities and Semantic Search

Why matching words isn't enough, how search moved from strings to things, and what entity salience can and can't tell you about your content.

Teaching series · Lesson 5 · beginner to intermediate

SEARCH: "UNMANNED AERIAL VEHICLE"

KEYWORD MATCH	
unmanned aerial vehicle	✓
drone	✗
UAV	✗
quadcopter	✗
Drohne (German)	✗

MEANING (ENTITY + EMBEDDING) MATCH	
unmanned aerial vehicle	✓
drone	✓
UAV	✓
quadcopter	✓
Drohne (German)	✓

Same thing, five strings. Keyword search finds one. Illustrative.

45	7	1
minutes	learning questions	entity analysis exercise

Ameneh Saeednia
Senior SEO and Generative Search Strategist

Before you start

What this lesson covers

My January 2024 lesson draft opened with seven questions. They're still the right ones, so this lesson answers them, with a few corrections.

By the end of this lesson you can

- Explain the **three problems** with pure keyword search.
- Define an **entity**, a **knowledge graph** and a **semantic network**.
- Describe how **embeddings** make meaning-based and cross-language search possible.
- Say what an **entity salience score** measures, and what it doesn't.
- Write content that makes the **main entity clear** and well connected.

Before this lesson

Lesson 1 introduces **vectors**; Lesson 2 introduces **triples and knowledge graphs**. Both help, neither is required.

How to use it

Read Parts 1 to 5, run the **entity analysis exercise on page 8**, then take the **quiz on page 9**.

Part 1

Three problems with keywords

Keyword search matches **strings of characters**. People describe the same thing in many ways, and one word can mean many things.

1

Too little

Synonyms, misspellings and phrasing variants **don't match**. “Incident,” “event” and “problem” may all describe one thing.

2

Too much

One word, many meanings. **“Interest”** can mean curiosity, a loan charge or a legal stake. Without context, matches are noisy.

3

Unknown words

Searchers often **don't know** the right term until they've read the material. Experts and beginners use different words.

A classic result

In a 1985 study of a legal document system, lawyers believed their keyword searches had found about **75%** of relevant documents. Blair and Maron measured it at about **20%**. [1] The searches were careful; the vocabulary problem was the limit.

Part 2

From strings to things

In 2012 Google introduced its Knowledge Graph with the phrase **“things, not strings.”** [2] The idea: understand the real-world entity a word refers to.

Term	Meaning	Example
Entity	A uniquely identifiable thing: person, place, organization, concept, event	The Colosseum (the monument, not the word)
Knowledge graph	A network of entities and the relationships between them	Colosseum → located in → Rome
Semantic network	An older term for the same idea: concepts as nodes, relations as links	Often used interchangeably
Co-occurrence	Entities that regularly appear together in text and queries	Colosseum, Roman Forum, Palatine Hill

Why disambiguation matters

“Paris” could be the French capital, a city in Texas or a person. Search systems use **surrounding entities** to decide. “Paris, the Eiffel Tower and the Seine” leaves no doubt.

Where trust comes in

Entities that are **consistently described** across your site, other sites and structured data are easier to identify confidently. That's the link to the *Entity Clarity* playbook.

Part 3

How meaning-based search works

Semantic search uses **embeddings**: each word, passage or document becomes a vector, and texts with similar meanings sit close together.

1

Distance is similarity

A search for “alcoholic drink” can match **“cocktail”** or **“margarita”** because their vectors are close, even with no words in common.

2

Near-duplicates show up

Two pages that say the same thing in different words have **very close vectors**. That's useful for finding overlap on your own site.

3

Languages can align

Multilingual models place **“drone,” “UAV” and “Drohne”** near each other, so meaning can be matched across languages.

4

Keywords still help

Exact names, codes and rare terms are matched best **lexically**. Modern systems usually combine both: hybrid retrieval.

Is semantic search language-dependent?

The model is trained on languages, so quality varies by language. But **meaning can be matched across languages** when the model is multilingual. That answers one of the 2024 draft's questions.

Part 4

Entity salience: useful, and often overstated

Google Cloud's Natural Language API returns entities in a text with a **salience score from 0 to 1**: how central each entity is to that text. [3]

Common claim (including my 2024 draft)	More accurate version
"Google's salience score"	It's the **Cloud Natural Language API's** score. Google Search's internal signals aren't public.
High salience means the page will rank	It means **this API** sees the entity as central to the text. Ranking depends on much more.
Salience of a keyword	Salience is measured for **entities**, not keyword strings.
High salience shows well-structured content	It shows the entity is **prominent**. Structure and quality are separate questions.

What it's good for

A **clarity check**: is the entity you think the page is about the one an NLP system finds most central? If the top entity is your brand or a side topic, the text may be unclear.

What moves it

Naming the entity early, making it the **subject of key sentences**, and connecting it to related entities. Repeating it as a keyword helps far less than you'd think.

Part 5

Writing for entity-based search

Five habits that make your main entity easier for people and machines to identify.

1

Name it plainly and early

Use the entity's **full, common name** in the title, first sentence and main heading.

2

Disambiguate

Add the context that rules out other meanings: **location, category, time period**.

3

Cover related entities

Mention the people, places and concepts **genuinely connected** to it. That's what co-occurrence reflects.

4

State relationships as triples

“[Entity] is located in..., was founded by..., is used for...” (Lesson 2).

5

Link identities

Use **sameAs** and consistent profiles so your entity connects to its record elsewhere.

Where this connects

My posts *Entity SEO 101* and *Named Entity Recognition: the SEO Basic Almost Nobody Actually Checks* go deeper on both topics.

Exercise

Run an entity analysis on one page

Time: 15 minutes. Use the **Google Cloud Natural Language API demo** as a proxy, not as a view of Google Search.

Step	What to do
1	Copy the main body text of one page. Paste it into the demo and choose **Entities**.
2	Write down the **top five entities** and their salience scores.
3	Is the entity the page is meant to be about **first**? Is it clearly ahead of the second?
4	Check for **wrong entity types** or misidentified entities, such as a place read as a person.
5	Rewrite the **opening two paragraphs** using the five habits from Part 5. Run it again.
6	Record both results and note **what changed**. Don't treat the score as a ranking signal.

What good looks like

The main entity is **first and clearly ahead**, correctly typed, with related entities appearing below it.

Discussion question

If your brand name outranks your topic, is the page **about your brand** or about the topic? Which should it be?

Check your understanding

Quick quiz

Answers are on the next page.

1. Which is a problem with pure keyword search?

a) It finds every synonym
b) It misses different words for the same thing
c) It understands context too well

2. “Things, not strings” refers to:

a) Matching entities, not just words
b) Removing stop words
c) Counting keywords

3. Semantic search finds “cocktail” for “alcoholic drink” because:

a) The words match exactly
b) Their embeddings are close in meaning
c) Of keyword density

4. An entity salience score of 0.8 from the Natural Language API means:

a) The page will rank highly
b) The API sees that entity as central to the text
c) Google Search gave it 0.8

5. The best way to make the main entity clear is:

a) Repeat it in every sentence
b) Name it early, disambiguate it and connect related entities
c) Hide it in the footer

Answers and sources

Answer key, glossary and further reading

Answer key

1: **b**.
2: **a**.
3: **b**.
4: **b**. A property of that API, not of Search.
5: **b**.

Glossary

Entity: a uniquely identifiable thing.
Knowledge graph: entities connected by relationships.
Embedding: text turned into a vector.
Salience: how central an entity is to a text.
Disambiguation: deciding which meaning is intended.
Co-occurrence: entities appearing together.

01 **An evaluation of retrieval effectiveness for a full-text document-retrieval system. Blair and Maron, Communications of the ACM, 1985**
The 75% estimated vs. about 20% measured recall result.
https://doi.org/10.1145/3166.3197

02 **Introducing the Knowledge Graph: things, not strings. Google, May 2012**
https://blog.google/products/search/introducing-knowledge-graph-things-not/

03 **Analyzing entities. Google Cloud Natural Language API**
Entity types and salience scores used in the exercise.
https://cloud.google.com/natural-language/docs/analyzing-entities

Lesson by Ameneh Saeednia, built on my January 2024 lesson draft and prepared with AI assistance. Examples are illustrative.