NERD: Named Entity Recognition and Disambiguation

How machines find the things your page mentions, decide which real-world thing each one is, and link it to a knowledge base. And how to write so they get it right.

Teaching series · Lesson 11 · intermediate

ONE SENTENCE, FOUR STEPS

"Jaguars are the fastest feline in the Pantanal."

1 Recognize
Find the mentions: **Jaguars**, **Pantanal**.

→

2 Candidates
Jaguar the cat · Jaguar Cars · Jacksonville Jaguars

→

3 Disambiguate
feline and **Pantanal** point to the cat.

→

4 Link
Jaguar → the animal's Wikipedia / Knowledge Graph ID

4	8	2023
steps in the pipeline	writing rules	draft, corrected

Ameneh Saeednia
Senior SEO and Generative Search Strategist

Before you start

What this lesson covers

Search engines and AI systems don't read words, they read **things**. NERD is the step that turns a word on your page into a specific thing in a knowledge base. This lesson is based on a **2023 draft from my archive**, checked against Google's documentation and corrected where it was wrong.

By the end of this lesson you can

- Explain the three subtasks of NERD: **recognition, candidate generation and disambiguation**.
- Describe the **signals** a system uses to pick the right entity.
- Write sentences that make the right entity **easy to recognize**.
- Use **facts, research tools and markup** to confirm which entity you mean.
- Spot the **outdated claims** in older entity-SEO advice.

Related lessons

Lesson 5 (entities and semantic search, salience), **Lesson 7** (entity schema: about and mentions) and **Lesson 8** (knowledge graphs). This lesson sits between them: how a mention becomes an entity.

Honest framing

Google doesn't publish how its own NERD systems work. The pipeline here is the **standard one in NLP research**, and Google's public statements fit it. Where something is inference, the lesson says so.

Part 1

What NERD means

NERD stands for **Named Entity Recognition and Disambiguation**. In NLP it's usually called **entity linking**: assigning a unique identity to the people, places and organizations mentioned in a text. [1]

Subtask	Question it answers	Example: “Paris is the capital of France.”
1 · Named entity recognition	Which words are names of things?	“Paris” and “France” are named entities.
2 · Candidate generation	Which known entities could this name refer to?	Paris the city, Paris Hilton, Paris in Texas, Paris the mythological prince...
3 · Disambiguation	Which candidate is meant here?	“capital” and “France” point to **Paris, the city in France**.

Google's own example

When Google launched the Knowledge Graph in 2012, it explained the problem with **Taj Mahal**: do you mean the monument, or Taj Mahal the musician? It called this finding **the right thing**. [2]

The knowledge base behind it

Google said the Knowledge Graph drew on public sources such as **Freebase, Wikipedia and the CIA World Factbook**, and held more than **500 million objects and 3.5 billion facts** at launch. [2]

In one line

Recognition finds the **mention**. Disambiguation picks the **entity**. Linking gives it an **ID**, such as a Wikipedia URL or a Knowledge Graph MID.

Part 2

The pipeline, step by step

Here is the pipeline on the 2023 draft's best example: **jaguar**, which can be an animal, a car brand or a football team.

"Jaguars are the fastest feline in the Pantanal."

1 Recognize
Find the mentions: **Jaguars**, **Pantanal**.

→

2 Candidates
Jaguar the cat · Jaguar Cars · Jacksonville Jaguars

→

3 Disambiguate
feline and **Pantanal** point to the cat.

→

4 Link
Jaguar → the animal's Wikipedia / Knowledge Graph ID

Step	What the system looks at	What you control
Recognize	Capitalization, grammar, known names, entity types (person, location, organization, event, product…)	Use the **full proper name** at least once, spelled the standard way
Candidates	Every knowledge-base entry whose name or alias matches the mention	Nothing directly. Know which **namesakes** compete with your entity
Disambiguate	Context words, other entities nearby, the page's topic, how common each meaning is	**Context**: type words, related entities, facts
Link	The chosen entity's ID in the knowledge base	**Markup** with `sameAs` to the right Wikipedia or Wikidata page

What Google's API shows

Google's Cloud Natural Language API (v1) returns, for most entities, a **Wikipedia URL** and a **Knowledge Graph MID** when one is found. [3] That is entity linking you can see.

Part 3

How disambiguation decides

Entity-linking systems weigh several kinds of evidence. Exact methods vary by system, but these signals are standard in the research.

1

Context words

Words around the mention: **feline, prey, rainforest** point to the cat; **sedan, engine, dealer** to the car.

2

Co-occurring entities

Other entities on the page should **fit together**. Pantanal and Brazil fit the cat. Land Rover fits the car.

3

Page topic

The **title, headings and main subject** of the page shape every mention below them.

4

Popularity

With little context, systems lean on the **most common meaning**. A small brand that shares a famous name starts behind.

5

Type fit

If the sentence needs a **place** or a **person**, candidates of the wrong type drop out.

6

Explicit links

Links and **structured data** that point at a specific entity page remove most of the doubt.

Candidate for “Jaguars”	Evidence for	Evidence against	Result
Jaguar (the cat)	“feline”, “Pantanal”	None	**Linked**
Jaguar Cars	“fastest”	“feline”, a wetland region	Rejected
Jacksonville Jaguars	Plural form	No sport, league or city	Rejected

Part 4

Write so the right entity is recognized

Most of the time, disambiguation happens naturally when you write clearly. These rules matter most when your entity **shares a name** with something more famous.

Jaguar sightings are up this year.	**Jaguar** (*Panthera onca*) sightings in the **Pantanal** rose this year. Type cue + location entity
Apple announced new features.	**Apple Inc.** announced new features for the **iPhone**. Full name + related product
Our founder **Sam** started the company in 2019.	**Sam Rivera**, founder of **[Company]**, started it in **2019** in **Vancouver**. Full name + role + date + place
Like **Tom Cruise**, our guides do their own stunts.	Our guides lead **hiking routes** above 2,000 m, with ropes on steep sections. No competing celebrity entity

1

Full name first

Use the **complete name** on first mention, short forms after.

2

Add a type cue

“the big cat”, “the car maker”, “the city in Texas”.

3

Bring its neighbours

Mention the entities that **define** it: places, founders, products.

4

Be consistent

One spelling, one name. Variants create **new candidates**.

About jokes and references

The 2023 draft said **never** use analogies. That's too strong. The real risk is a reference that adds a **competing entity** (a celebrity, a film) near your main topic. If you use one, keep it short and clearly separate.

Part 5

Facts confirm the entity

Facts that only one candidate can match are the strongest disambiguation signal. The 2023 draft had a good idea here, but its facts were **undated and partly stale**.

2023 draft: mixed years, no sources

```
Did you know Paris is the capital of France? ... it
has 2,148,271 people living within its borders as of
2020. It takes up 105 square kilometers... It had a
GDP of $808 billion in 2017... serves 5.23 million
passengers daily.
```

Better: stable facts plus a dated, sourced figure

```
Paris is the capital of France and its most populous
city. The city covers about 105 km² (41 sq mi). Its
population was [figure] in [year], according to
INSEE. The Paris Métro opened in 1900.
```

Why it works

Capital of France, 105 km², Métro 1900: only one Paris matches all of these.
Each fact is also a clean **triple** (Lesson 10).
A figure that changes is given a **year and a source**.

The risk

A wrong or out-of-date number is a **wrong fact about the entity**. Google research has explored judging sources by the accuracy of the facts they state (Lesson 10). **Date it, source it, or leave it out.**

Part 6

Research the entities for a page

Before writing, list the main entity, its competing namesakes, and the related entities a complete page should mention.

Tool	What to look for	Watch out for
Google Search	Knowledge panel, **People also search for**, People also ask, related searches	Results are personalised by location and history
Wikipedia	Lead section links, infobox fields, **disambiguation pages** (e.g. “Jaguar (disambiguation)”)	Not every entity has an article. Many entities in Google's graph don't
Wikidata	The item's **properties**: instance of, country, founder, inception	Community-edited. Check important values
Google NL API (v1)	Which entities it finds, their **type**, **salience**, Wikipedia URL and MID [3]	The newer **v2 API does not return salience** [4]
Top-ranking pages	Entities competitors mention that you don't	Don't copy. Look for **gaps in coverage**

Your entity list

Main entity, with its Wikipedia or Wikidata URL.
Namesakes to steer away from.
5 to 10 related entities that define it.
3 to 5 facts only the right entity matches.

Correction

The 2023 draft said Google had **discontinued** entity salience. Not quite: the **v1** API still returns salience, the **v2** API doesn't. [3][4]

Part 7

Markup that removes doubt

Text gives context. Structured data can name the **exact entity** with a link to its knowledge-base page. This is the **linking** step, done for the machine.

Article about the jaguar, with about and mentions

```
{
  "@context": "https://schema.org",
  "@type": "Article",
  "@id": "https://example.com/jaguar-spotting-
pantanal/#article",
  "headline": "Where to See Jaguars in the Pantanal",
  "about": {
    "@type": "Taxon",
    "name": "Jaguar",
    "sameAs": "https://en.wikipedia.org/wiki/Jaguar"
  },
  "mentions": [
    {"@type": "Place", "name": "Pantanal",
     "sameAs":
"https://en.wikipedia.org/wiki/Pantanal"},
    {"@type": "Country", "name": "Brazil",
     "sameAs":
"https://en.wikipedia.org/wiki/Brazil"}
  ]
}
```

What each part does

`about` names the **main entity**: the animal, not the car.
`sameAs` points to the **Wikipedia page** for that exact entity.
`mentions` lists **supporting entities** that also help disambiguation.
The markup must match what the **visible text** says. [5]

Common mistakes

Putting a Wikipedia `sameAs` on your **Organization** when it describes the topic, not you.
Linking a **disambiguation page** instead of the entity's own article.
Listing entities in `mentions` that the page **never mentions**.

Go deeper

Lesson 7 covers `about`, `mentions` and `sameAs` in detail, including how to choose between Wikipedia and Wikidata URLs.

Part 8

What the 2023 draft got wrong

The draft's core idea holds up: **help machines recognize the right entity with context and facts**. Several details needed correcting.

2023 claim	What's accurate
"Google uses Wikidata and Wikimedia's database."	Inference. Google named **Freebase, Wikipedia and the CIA World Factbook** in 2012, plus much larger-scale sources. [2]
Entities are things that have a Wikipedia page, basically.	No. Knowledge graphs hold far more entities than Wikipedia has articles. A Wikipedia page **helps**, it doesn't **define** an entity.
"The named entity is your main keyword, basically."	A keyword is a **string**; an entity is a **thing** with an ID. Several keywords can point to one entity.
"Don't use analogies and jokes."	Too strong. Avoid references that add **competing entities**.
Entities will soon replace backlinks and keywords.	Speculation. Google hasn't said this.
"Claim your GMB profile."	Google My Business is now **Google Business Profile**.
Google discontinued entity salience.	**v1** still returns salience; **v2** doesn't. [3][4]
Paris facts with mixed years	Stable facts are great; changing figures need a **year and source**.

Exercise

A NERD check for one page

Time: 30 minutes. Pick a page whose main entity shares a name with something else, or a page about your own brand.

Step	What to do
1	Write down the page's **main entity** and find its **Wikipedia or Wikidata URL**. If none exists, note its official site.
2	List the **namesakes** that compete for that name. Check Wikipedia's disambiguation page.
3	Read the **first 150 words**. Is the full name there? Is there a **type cue**? Which **related entities** appear?
4	Add **3 facts** only the right entity matches. Date and source anything that changes.
5	Remove references that bring in a **competing entity**.
6	Paste the text into Google's **Natural Language demo**. Does it find the right entity and link the right Wikipedia page?
7	Add or fix `about`, `mentions` and `sameAs` in the page's JSON-LD.

What good looks like

The NL demo links the **right** Wikipedia page for the main entity. The first paragraph names it fully, with a type cue and **at least two** defining entities.

Discussion

Which namesake was hardest to steer away from? What **single fact** did the most to settle it?

Check your understanding

Quick quiz

1. Entity linking is made of which three subtasks?

a) Crawling, indexing, ranking
b) Recognition, candidate generation, disambiguation
c) Keywords, links, schema

2. In “Jaguars are the fastest feline”, which word settles the meaning?

a) fastest
b) are
c) feline

3. For most entities, Google's NL API v1 returns:

a) A Wikipedia URL and a Knowledge Graph MID
b) A ranking score
c) The page's backlinks

4. Which fact best confirms you mean Paris, France?

a) Paris is popular
b) Paris covers about 105 km² and its Métro opened in 1900
c) Paris has many hotels

5. Which statement is accurate?

a) Only things with a Wikipedia page are entities
b) A keyword and an entity are the same
c) Several keywords can refer to one entity

6. In JSON-LD, which property names the page's main entity?

a) mentions
b) about
c) author

Answer key and sources

Answer key

1 b. NER, candidates, disambiguation. **2 c.** Only the cat is a feline. **3 a.** Wikipedia URL and MID, when available. **4 b.** Facts only one Paris matches. **5 c.** A keyword is a string; an entity is a thing. **6 b.** `about` names the main entity; `mentions` lists supporting ones.

01 **Entity linking. Wikipedia**
Definition of entity linking (also called NERD) and its three subtasks.
https://en.wikipedia.org/wiki/Entity_linking

02 **Introducing the Knowledge Graph: things, not strings. Google, 2012**
Taj Mahal example; sources; 500M+ objects and 3.5B+ facts at launch.
https://blog.google/products/search/introducing-knowledge-graph-things-not/

03 **Entity (v1). Cloud Natural Language API reference**
Salience; metadata with wikipedia_url and mid.
https://cloud.google.com/natural-language/docs/reference/rest/v1/Entity

04 **Entity (v2). Cloud Natural Language API reference**
Fields: name, type, metadata, mentions, sentiment. No salience.
https://cloud.google.com/natural-language/docs/reference/rest/v2/Entity

05 **General structured data guidelines. Google Search Central**
Markup must represent the visible content.
https://developers.google.com/search/docs/appearance/structured-data/sd-policies

Lesson by Ameneh Saeednia, adapted from a 2023 draft in my archive, corrected against current documentation, and prepared with AI assistance. Examples are illustrative.