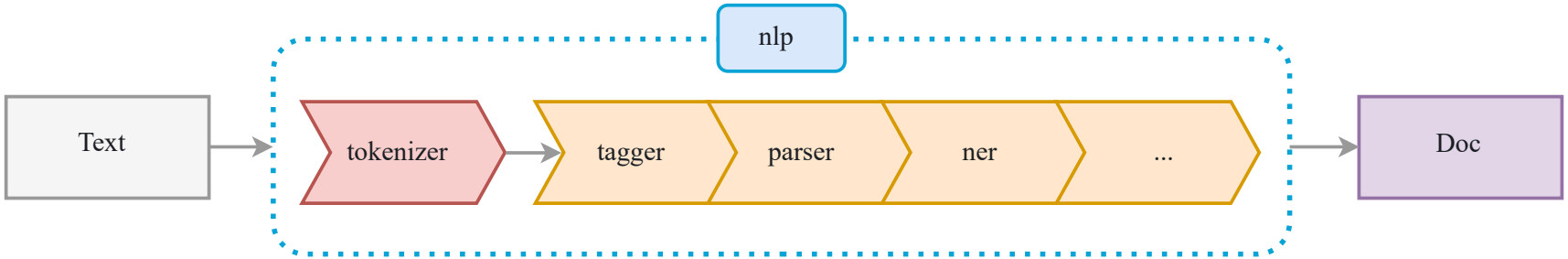
Entity linking functionality in spaCy:

Grounding textual mentions to knowledge base concepts

Sofie Van Landeghem
Freelancer ML and NLP @ OxyKodit

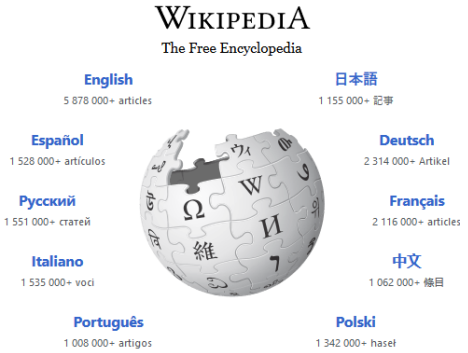
Entity Linking



The current spaCy *nlp* pipeline works purely on the textual information itself:

- Tokenizing input text into words & sentences
- Parsing syntax & grammar
- Recognising meaningful entities and their types
- …

But how can we ground that information into the “real world” (or its approximation – a knowledge base) … ?



Example of entity links

Who are all these Byron's in this text ?

Lord Byron PERSON expected his child to be a "glorious boy" and was disappointed when Lady Byron PERSON gave birth to a girl. The child was named after Byron PERSON 's half-sister, Augusta Leigh PERSON , and was called " Ada PERSON " by Byron PERSON himself.

Ada Lovelace

From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (*née* **Byron**; 10 l
general-purpose computer, the Analytical Engine. She was the
As a result, she is sometimes regarded as the first to recognise

Lovelace was the only legitimate child of the poet Lord Byron a
and left England forever four months later. He commemorated
War of Independence when Ada was eight years old. Her moth
Ada remained interested in Byron and was, upon her eventual
made Earl of Lovelace in 1838, Ada thereby becoming Counte

Lord Byron

From Wikipedia, the free encyclopedia

For the archaeologist, see George Byron Gordon (archaeologist).
"Byron" and "George Byron" redirect here. For other uses, see Byron (disam

George Gordon Byron, 6th Baron Byron FRS (22 January 1788 – 19 April 18
and is considered one of the historical leading figures of the Romantic movemen
are the lengthy narrative poems *Don Juan* and *Childe Harold's Pilgrimage*; many

He travelled extensively across Europe, especially in Italy, where he lived for sev
Shelley.[5] Later in life Byron joined the Greek War of Independence fighting the
at the age of 36 from a fever contracted after the First and Second Siege of Missolonghi.

Lady Byron

From Wikipedia, the free encyclopedia

This article is about Anne Byron, wife of Lord Byron. For the Austra

Anne Isabella Noel Byron, 11th Baroness Wentworth and Baron
and the wife of poet George Gordon Byron, more commonly known as

A highly educated and strictly religious woman, she seemed an unlikely
Beecher Stowe, revealed her fears about an alleged incest Lord Byron
where he had lived in 1810.

Their daughter Ada worked as a mathematician with Charles Babbage,

Complexity of the task

Synonymy

- Augusta Byron = Ada Byron = Countess of Lovelace = Ada Lovelace = Ada King

Polysemy

- 4 different barons were called “George Byron”
- “George Byron” is an American singer
- “George Byron Lyon-Fellowes” was the mayor of Ottawa in 1876
- …

Vagueness

- e.g. “The president”

Context is everything !

Some examples

Russ Cochran's PERSON reprints include The Complete EC Library ORG in black and white.

Russ Cochran: American golfer, or publisher ?

He felt that Rose PERSON and the Doctor's developing relationship was not subtle.

Rose: English footballer, or character from the TV series "Doctor Who" ?

This happened to DeLorean ORG owner Johnny Carson PERSON shortly after he was presented with the vehicle.

Johnny Carson: American talk show host, or American football player ?

NEL @ spaCy

Community feedback

spaCy @spacy_io · Feb 27

The v2.1 release, which focuses on stability and performance, is almost ready – so now it's time to start on some new features 🎁 Sofie (@OxyKodit) will be working on Named Entity Linking. If you know you need this, we'd love to hear your feedback! Thread: github.com/explosion/spaC...

Entity Linking in spaCy #3339

Open **svlandeg** opened this issue on 27 Feb · 38 comments

Feedback requested

We will start implementing the APIs soon, but we would love to hear your ideas, suggestions, requests with respect to this new functionality first!

36 5 21 7

- Cross-lingual mapping
- Link to Wikidata
- Train custom relationships / use your own KB
- ScispaCy (biomedical domain) as the perfect way to test our interfaces !

Design principles

For a first prototype, focus on WikiData instead of Wikipedia

- Stable IDs
- Higher coverage (WP:EN has 5.8M pages, WikiData has 55M entities)
- Better support for cross-lingual entity linking

Canonical knowledge base with potentially language-specific feature vectors

Do the KB reconciliation once, as an offline data-dependent step

In-memory (fast!) implementation of the KB, using a Cython backend

Processing Wikipedia

She married William King in 1835.

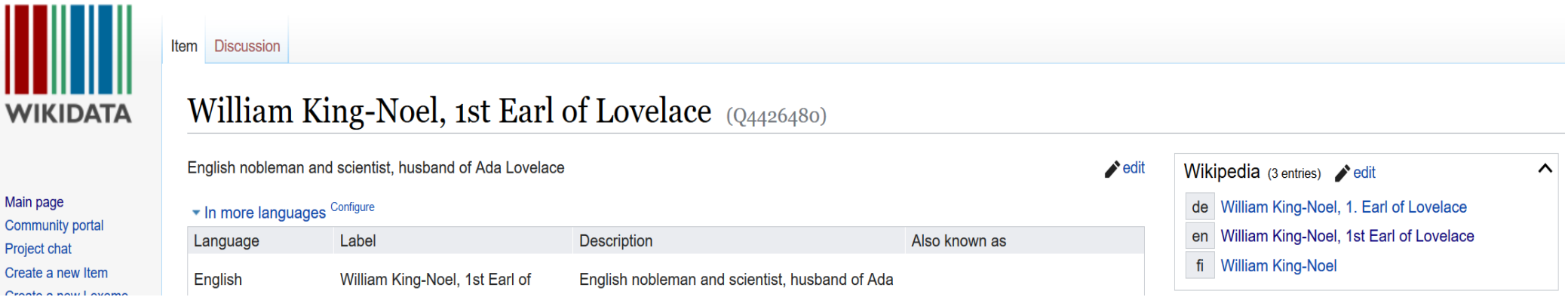
```
She married [[William King-Noel, 1st Earl of Lovelace|William King]] in 1835
```

Parsed aliases:

- 1st Earl of Lovelace
- Earl of Lovelace
- William King
- William King-Noel, 8th Baron King
- ...

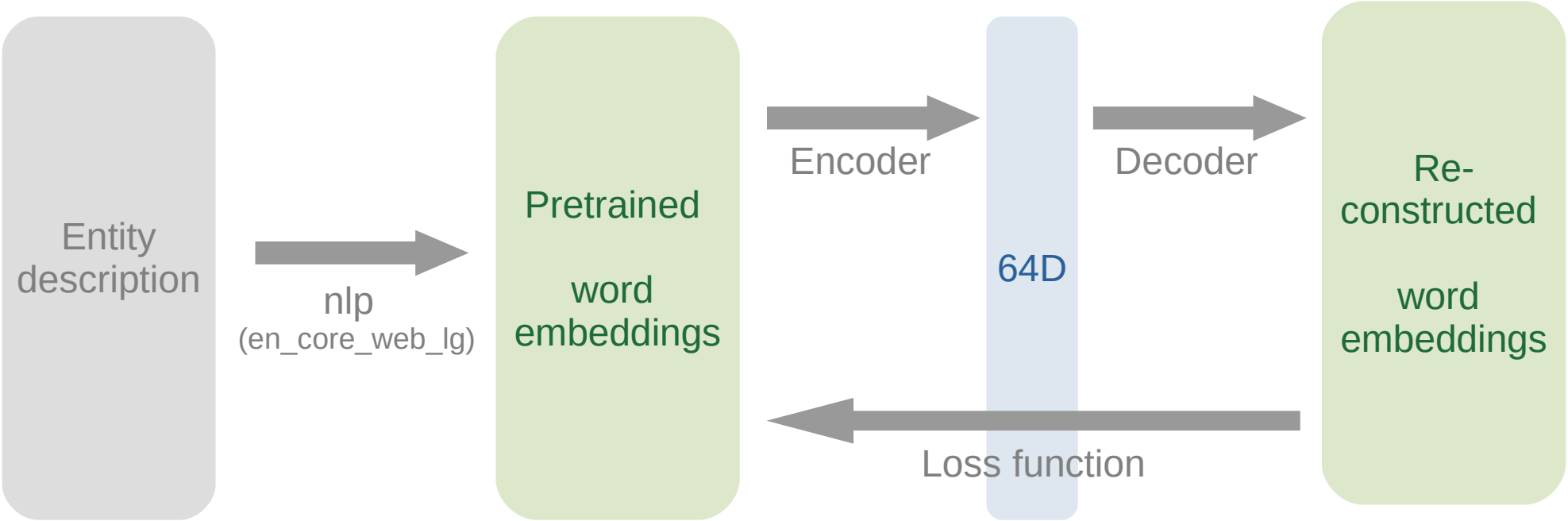
Aliases and prior probabilities from intrawiki links
Takes about 2 hours to parse 1100M lines of Wikipedia XML dump

Processing Wikidata

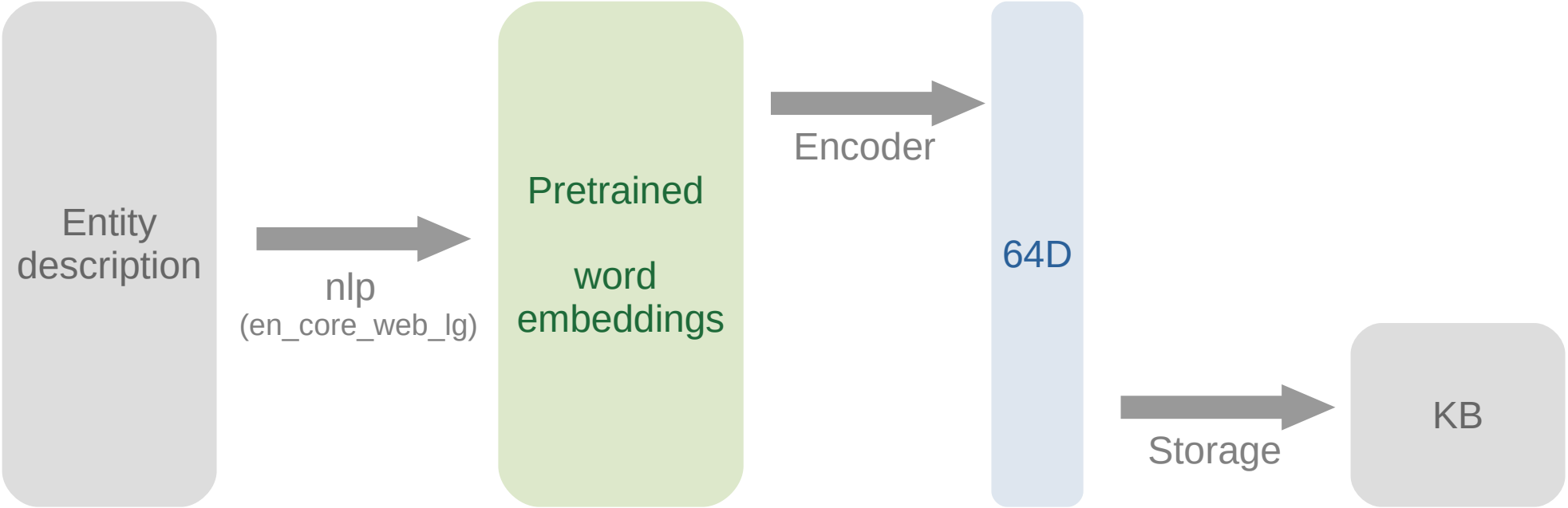


Takes about 7 hours to parse 55M lines of Wikidata JSON dump
→ Link English Wikipedia to interlingual Wikidata identifiers
→ Retrieve concise Wikidata descriptions for each entity

Entity encoder-decoder



Entity encoder



KB definition & storage

Some pruning to keep the KB manageable in memory:

- Keep only entities with min. 20 incoming interwiki links (from 8M to 1M entities)
- Each alias-entity pair should occur at least 5 times in WP
- Keep 10 candidate entities per alias/mention

KB size:

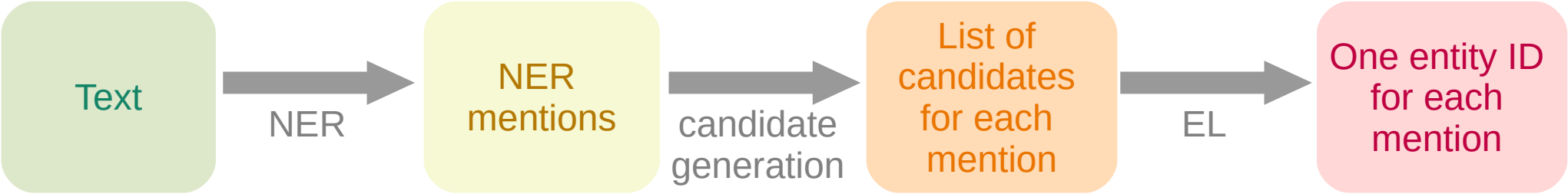
- ca. 1M entities and 1.5M aliases
- ca. 55MB file size without entity vectors
- ca. 350MB file size with 64D entity vectors
- Written to file, and read back in, in a matter of seconds

General flow

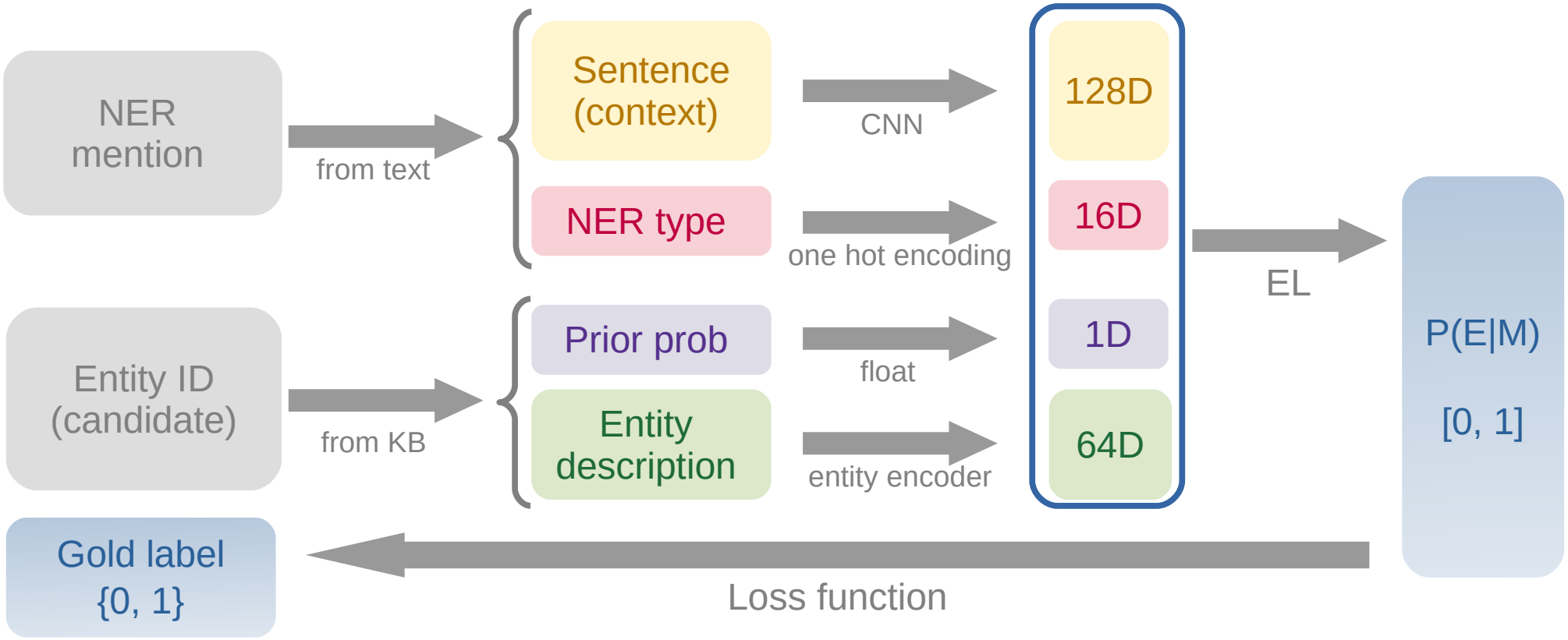
KB exposes functionality for **candidate generation**

- Input: An alias or textual mention (e.g. “Byron”)
- Output: list of candidates, i.e. (entity ID, prior probability) tuples
  - ➔ Currently implemented as the top X of entities, sorted by their prior probabilities

Within the list of candidates, the **entity linker** (EL) needs to find the best match (if any)



Entity linker



Code & results

Code examples

```
kb = KnowledgeBase(vocab=vocab, entity_vector_length=64)

kb.add_entity(entity="Q1004791", prob=0.2, entity_vector=v1)
kb.add_entity(entity="Q42", prob=0.8, entity_vector=v2)
kb.add_entity(entity="Q5301561", prob=0.1, entity_vector=v3)

kb.add_alias(alias="Douglas", entities=["Q1004791", "Q42", "Q5301561"], probabilities=[0.6, 0.1, 0.2])
kb.add_alias(alias="Douglas Adams", entities=["Q42"], probabilities=[0.9])
```

```
el_pipe = nlp.create_pipe(name='entity_linker', config={"context_width": 128})
el_pipe.set_kb(kb)
nlp.add_pipe(el_pipe, last=True)
```

```
other_pipes = [pipe for pipe in nlp.pipe_names
               if pipe != "entity_linker"]
with nlp.disable_pipes(*other_pipes):
    optimizer = nlp.begin_training()
    …
    nlp.update(…)
```

```
text = "Douglas Adams made up the stories as he wrote."
doc = nlp(text)

for ent in doc.ents:
    print(ent.text, ent.label_, ent.kb_id_)
```

Accuracy

Training data

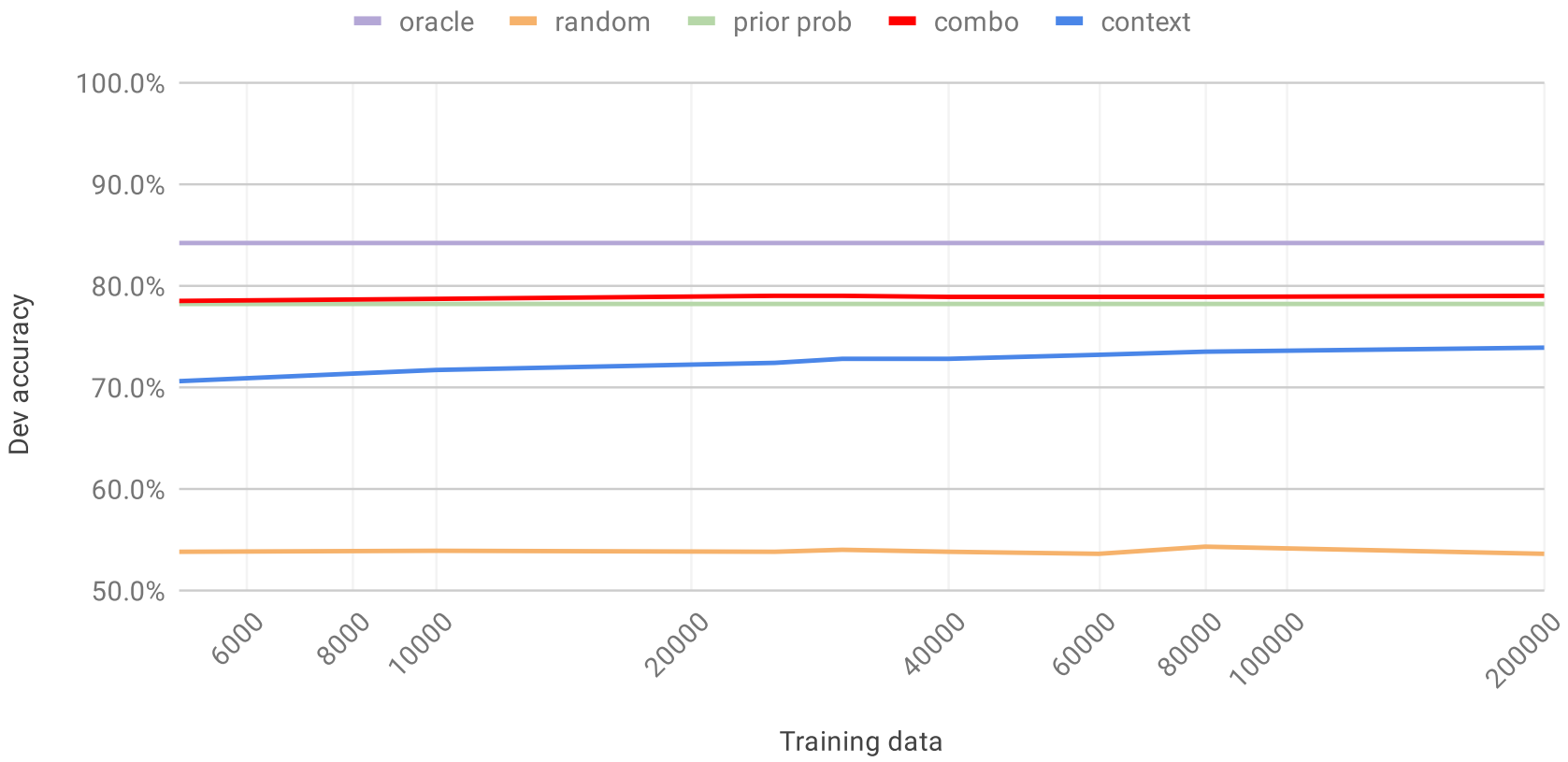
- Align WP intrawiki links with *en_core_web_lg* NER mentions
- Custom filtering: articles < 30K characters and sentences 5-100 tokens
- Trained on 200.000 mentions

KB has 1.1M entities (14% of all entities)

	Random baseline	Context only	Prior prob baseline	Context + prior prob	Oracle KB (max)
Accuracy %	54.0	73.9	78.2	79.0	84.2

The context encoder by itself is viable and significantly outperforms the random baseline.
It only marginally improves the prior prob. baseline though, and is limitated by the oracle performance.

Learning curve



Error analysis

Banteay Meanchey, Battambang, Kampong Cham, ... and Svay Rieng.
→ predicted “City in Cambodia” but should have been “Province of Cambodia”

Societies in the ancient civilizations of Greece and Rome preferred small families.
→ predicted “Greece” instead of “Ancient Greece”

Roman, Byzantine, Greek origin are amongst the more popular ancient coins collected
→ predicted “Ancient Rome” instead of “Roman currency” (but the latter has no description)

Agnes Maria of Andechs-Merania (died 1201) was a Queen of France.
→ predicted “kingdom in Western Europe from 987 to 1791” but should have been “republic with mainland in Europe and numerous oversea territories” (gold was incorrect)

Ongoing & future work

Define “a hill worth climbing”

- We need to obtain a better dataset that is not automatically created / biased
- Only then can we continue improving the ML models & architecture

Add in coreference resolution

- Entity linking for coreference chains (often not available in WP data)
- Improve document consistency of the predictions

Exploit the Wikidata knowledge graph

- Improve semantic similarity between the entities
- cf. OpenTapioca, Delpeuch 2019

Beyond Wikipedia & Wikidata:

- Reliable estimates of prior probabilities are more difficult to come by
- Candidate generation by featurizing entity names (e.g. scispaCy)

Thanks !