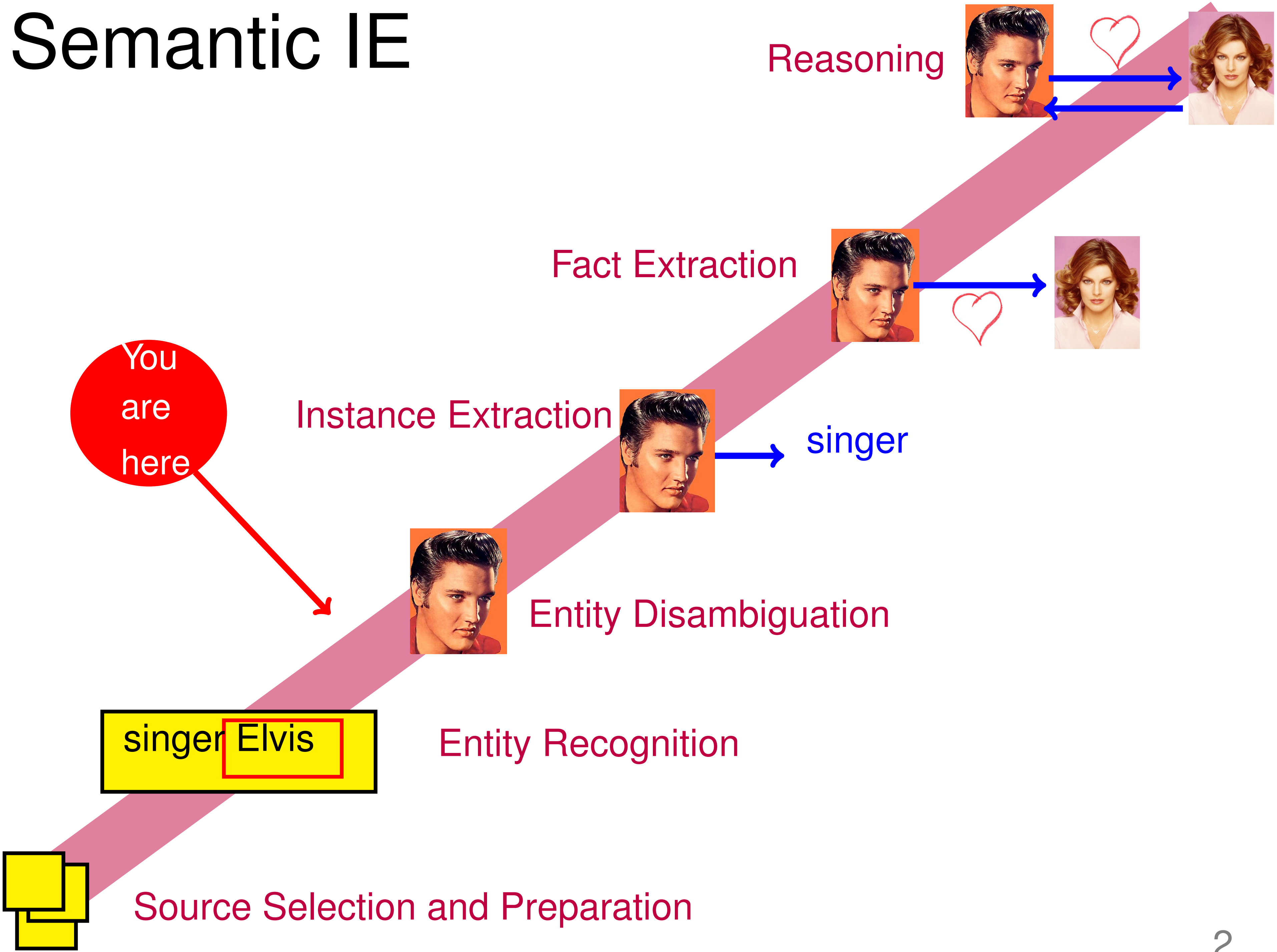
Disambiguation

Fabian M. Suchanek

Semantic IE



Reasoning

Fact Extraction

You are here

Instance Extraction

singer

Entity Disambiguation

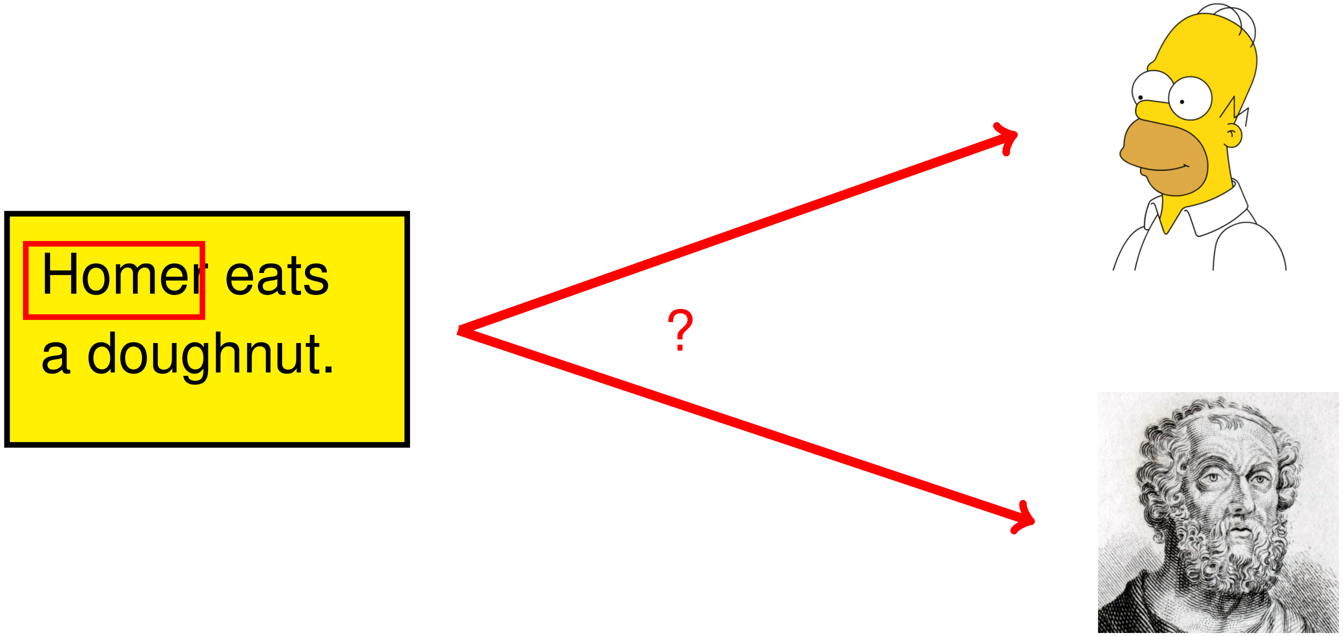
singer Elvis

Entity Recognition

Source Selection and Preparation

Def: Disambiguation

Given an ambiguous name in a corpus and its meanings, disambiguation is the task of determining the intended meaning.



Homer eats a doughnut.

?

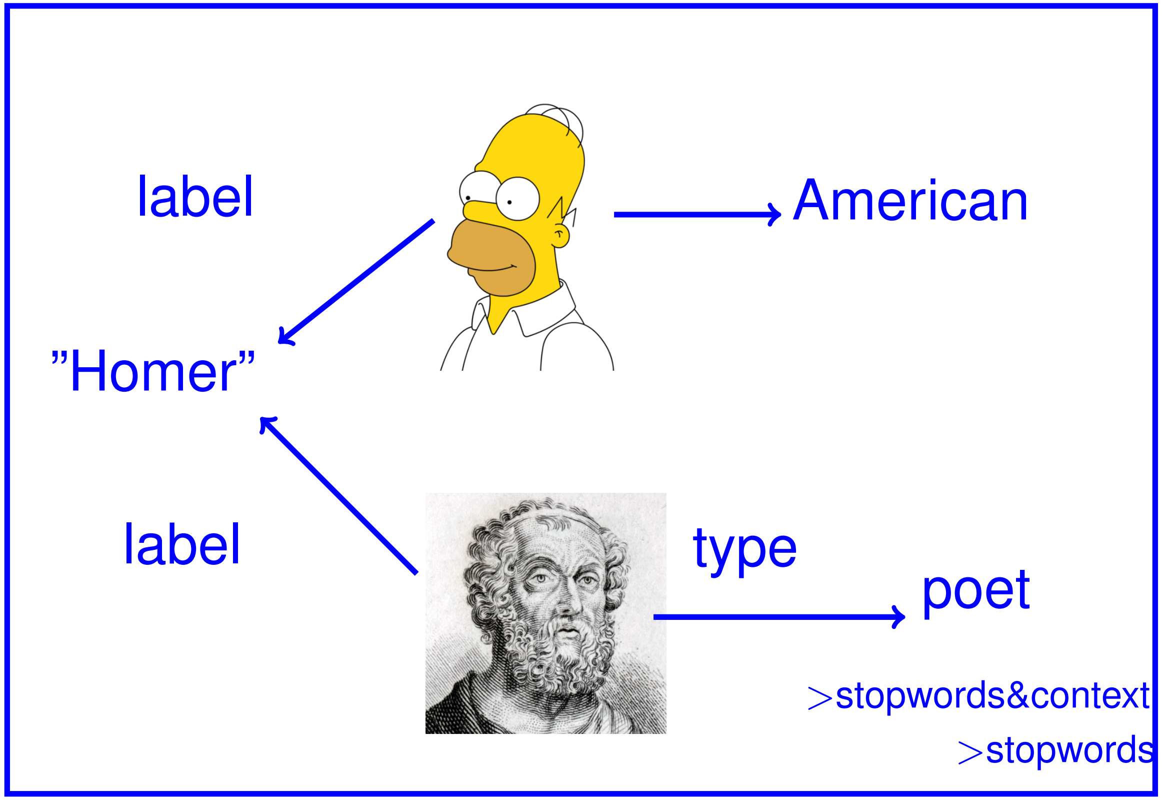
Disambiguation Setting

Usually Named Entity Recognition (NER) runs first, and the goal is to map the names to entities in a Knowledge Base (KB).

NER'ed
corpus

Homer eats
a doughnut.

Knowledge Base



label

"Homer"

label

American

type

poet

>stopwords&context

>stopwords

Def: Stopword

A stopword is a word that is common in a corpus but has little value for searching.

Definition from Postgres (The definition of stopwords is application-dependent)

Homer eats a doughnut.

Usually all words are stopwords, except

- nouns,
- adjectives
- non-auxiliary verbs

Example

Stopword Rationale

Imagine we search for

How many cats do the Simpsons have?

Here we do explain how many teeth the chicken have.

List of Simpson cats:
...

Stopword Rationale

Imagine we search for

How many cats do the Simpsons have?

Here we do explain how many teeth the chicken have.

Overlap: 5

List of Simpson cats:
...

Overlap: 2

Stopword Rationale

Imagine we search for

cats Simpsons?

Here we do explain how many teeth the chicken have.

Overlap: 0

List of Simpson cats:
...

Overlap: 2

task>50

Task: Stopwords

Remove the stopwords from the following sentences.

Don’t come here!
Homer was hit by Marge.
Homer ate a few doughnuts.

(These are fun examples where the meaning of the sentence changes. Usually, applications assume that the meaning of the sentence stay the same.)

Def: Context of a word

The context of a word in a corpus is the multi-set of the words in its vicinity without the stopwords.

(The definition may vary depending on the application)

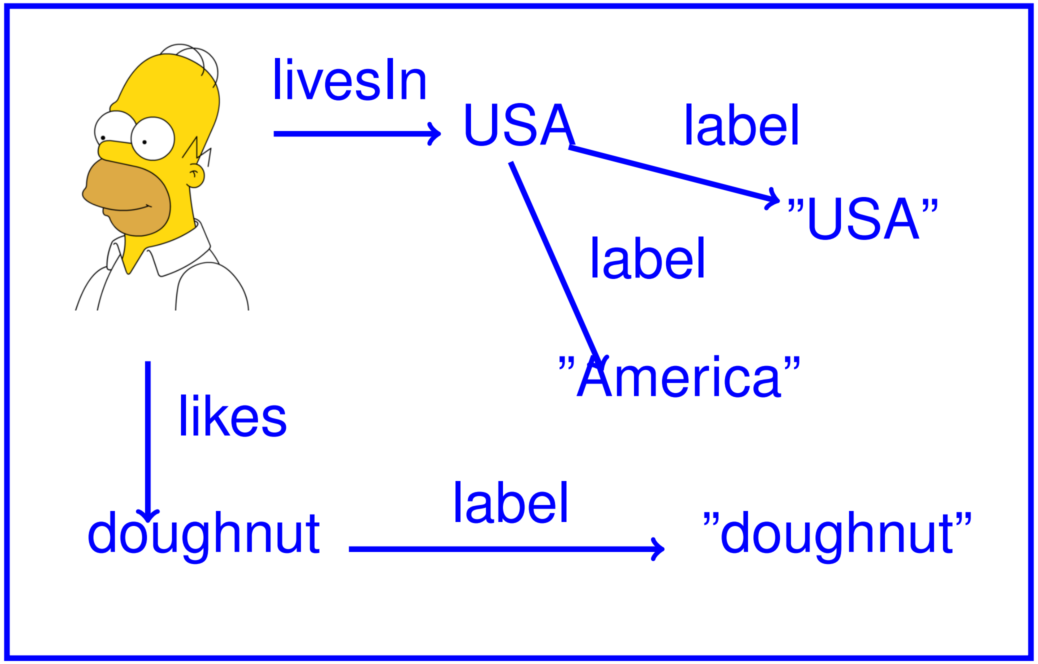
Homer eats a doughnut.

Context of “Homer”:
{eats, doughnut}

Def: Context of an entity

The context of an entity in a KB is the set of all labels of all entities in its vicinity.

(The definition may vary depending on the application)



livesIn

USA

label

"USA"

label

"America"

likes

doughnut

label

"doughnut"

Context of Homer:
{doughnut, USA,
America}

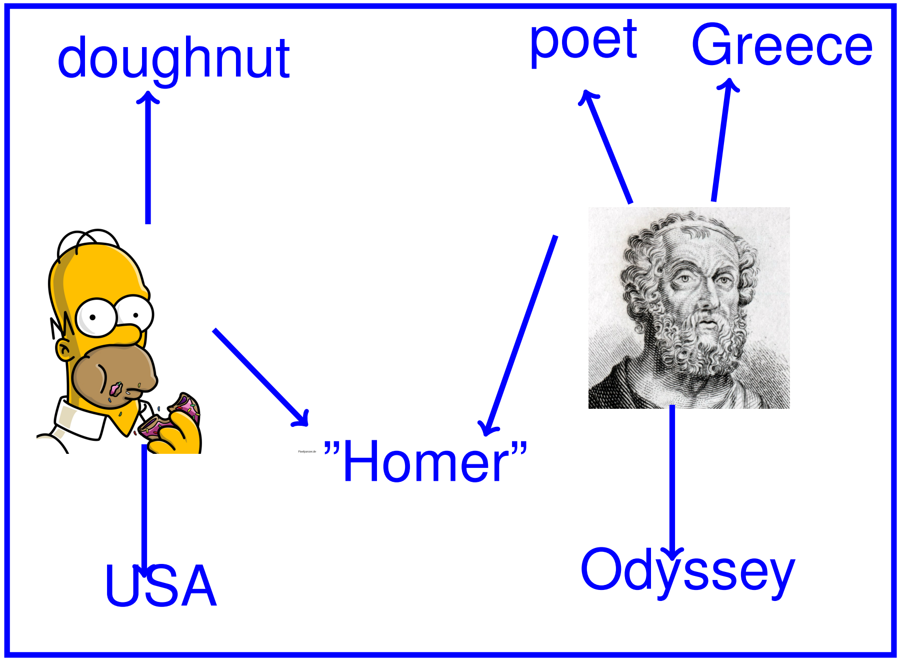
Def: Context-based disambiguation

Context-based disambiguation (also: bag of words disambiguation) maps a name in a corpus to the entity in the KB whose context has the highest overlap to the context of the name.

(The definition may vary depending on the application)

For USA Today, Homer is among the top 25 most influential people of the past 25 years.

Knowledge Base

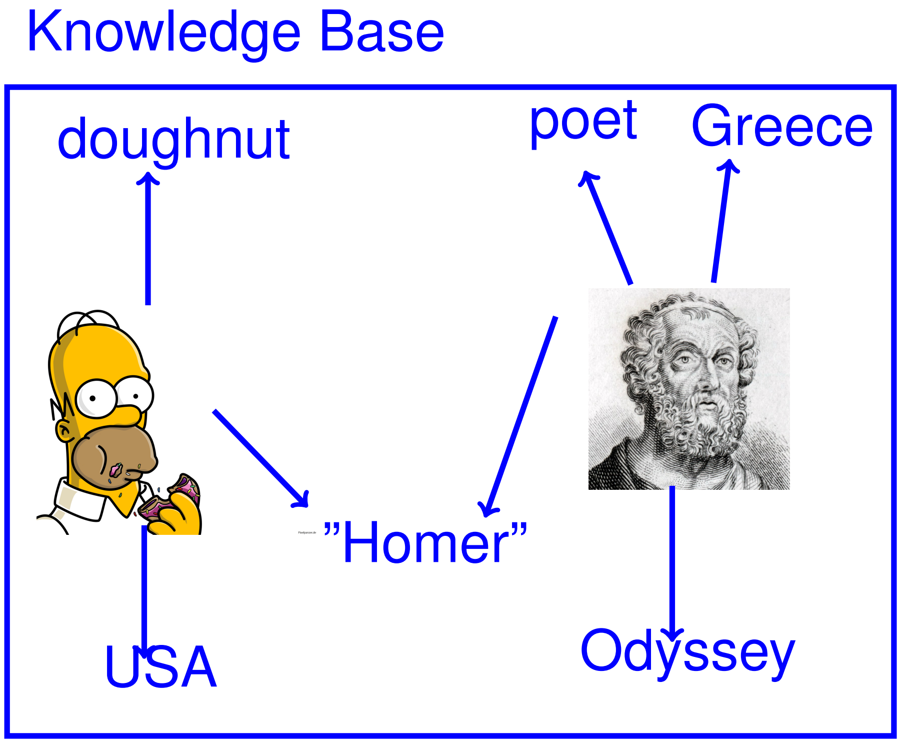


Def: Context-based disambiguation

Context-based disambiguation (also: bag of words disambiguation) maps a name in a corpus to the entity in the KB whose context has the highest overlap to the context of the name.

(The definition may vary depending on the application)

For USA Today, Homer is among the top 25 most influential people of the past 25 years.



Context of "Homer" in corpus: {USA, Today, top, influential, people, past, years}

Def: Context-based disambiguation

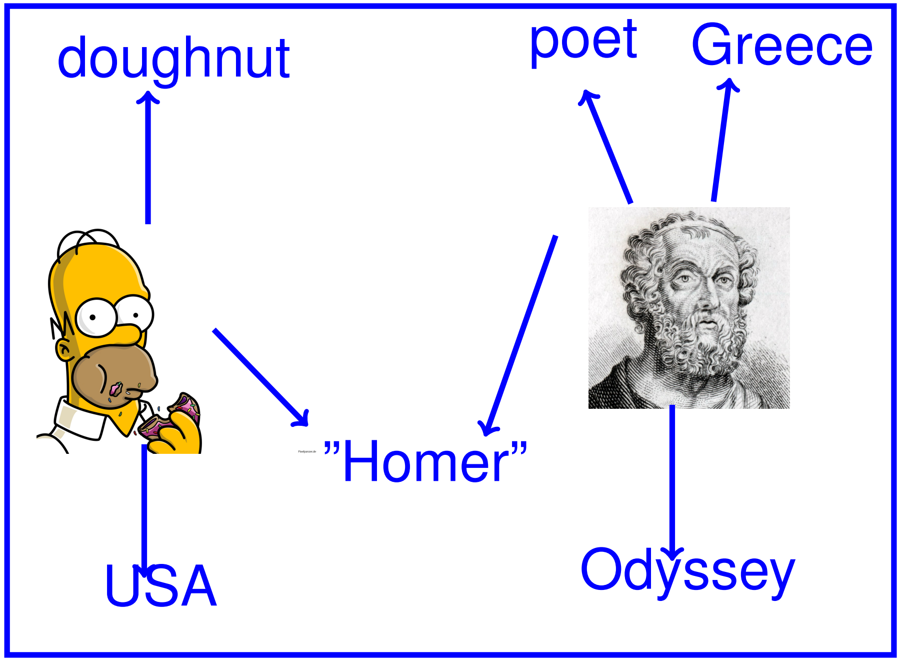
Context-based disambiguation (also: bag of words disambiguation) maps a name in a corpus to the entity in the KB whose context has the highest overlap to the context of the name.

(The definition may vary depending on the application)

For USA Today, Homer is among the top 25 most influential people of the past 25 years.

Context of ”Homer” in corpus: {USA, Today, top, influential, people, past, years}

Knowledge Base



{USA, doughnut} {poet, Geece, O.}

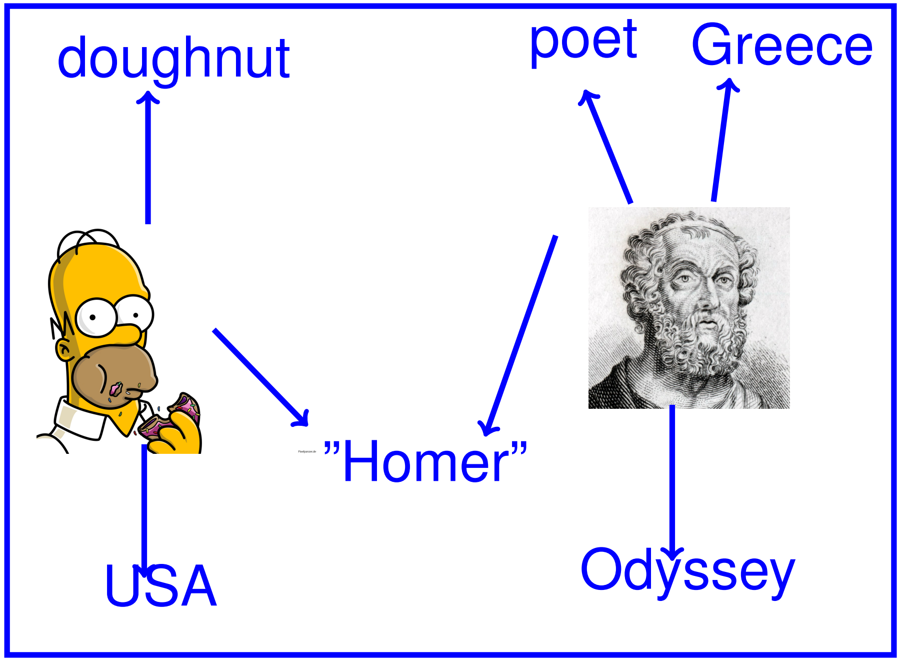
Def: Context-based disambiguation

Context-based disambiguation (also: bag of words disambiguation) maps a name in a corpus to the entity in the KB whose context has the highest overlap to the context of the name.

(The definition may vary depending on the application)

For USA Today, Homer is among the top 25 most influential people of the past 25 years.

Knowledge Base



doughnut

poet

Greece

"Homer"

USA

Odyssey

Context of "Homer" in corpus:
{USA, Today, top, influential, people, past, years}

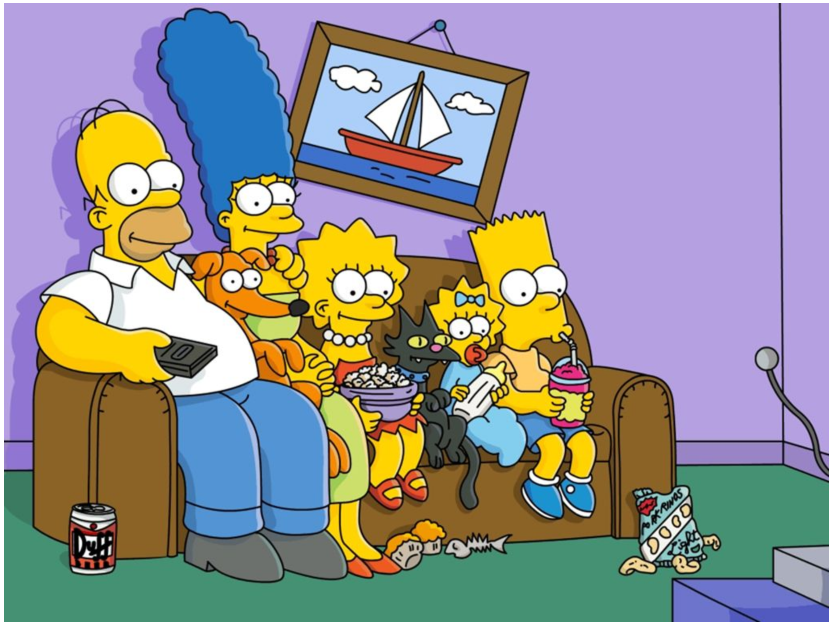
{USA, doughnut} {poet, Geece, O.}

highest overlap -> Homer Simpson wins

What if there is little context?

This is very important for the Simpsons.

?



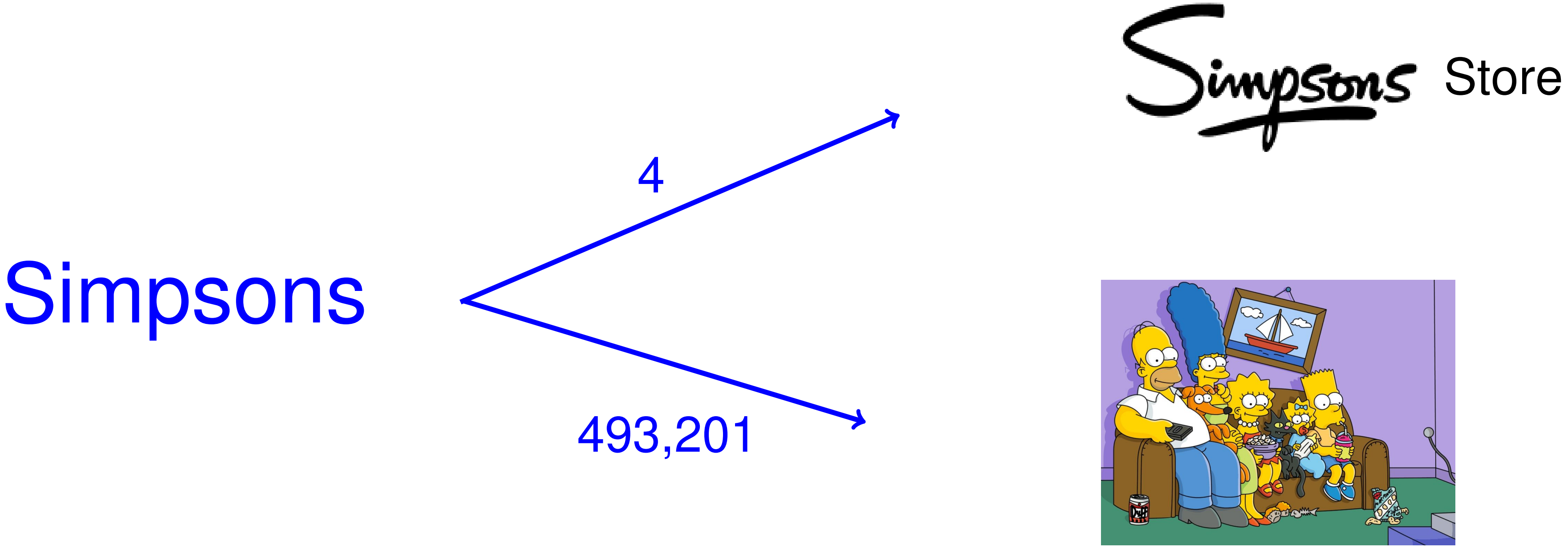
The Robert Simpson

Department Store.

Defunct since 1990.

Def: Disambiguation Prior

A disambiguation prior is a mapping from names to their meanings, weighted by the number of times that the name refers to the meaning in a reference corpus.



Simpsons

4

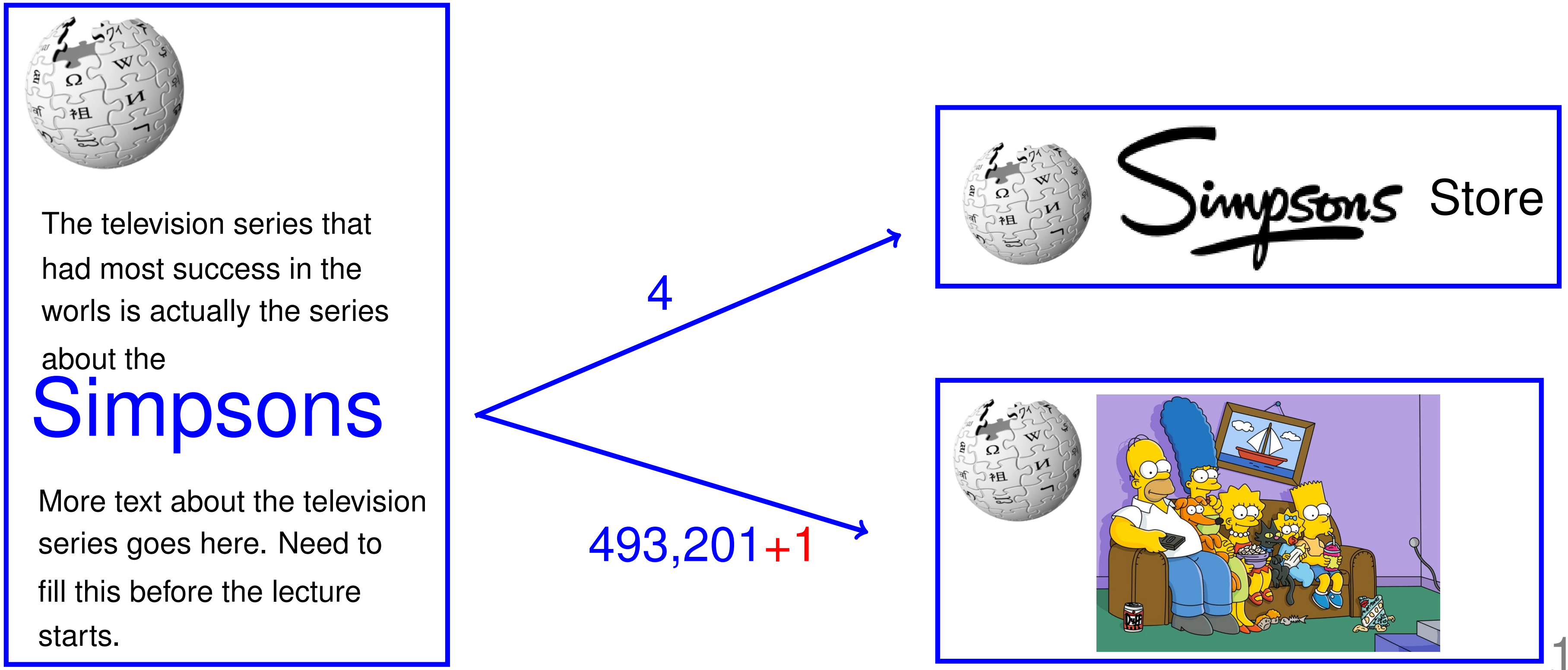
Simpsons Store

493,201

Def: Disambiguation Prior

A disambiguation prior is a mapping from names to their meanings, weighted by the number of times that the name refers to the meaning in a reference corpus.

It can be computed, e.g., from Wikipedia:



The television series that had most success in the worls is actually the series about the **Simpsons**

More text about the television series goes here. Need to fill this before the lecture starts.

4

Simpsons Store

493,201+1

Coherence

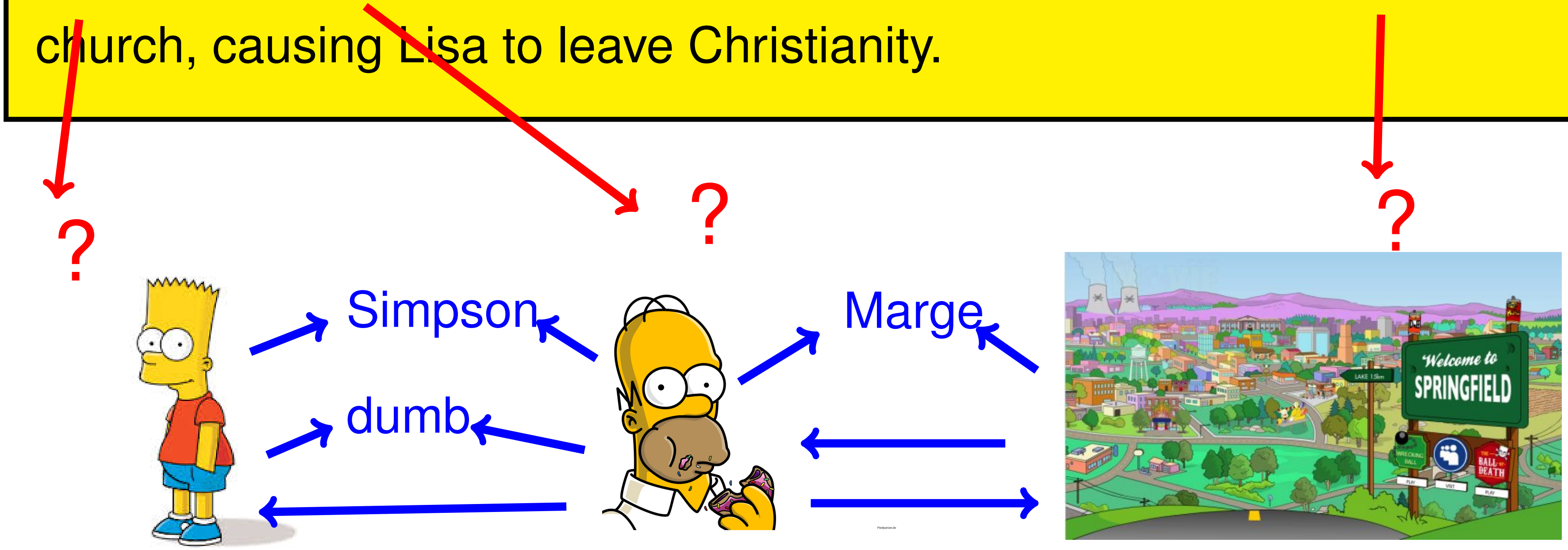
Bart and Homer accidentally launch a rocket into the Springfield church, causing Lisa to leave Christianity.

? ? ?

Welcome to
SPRINGFIELD

Coherence

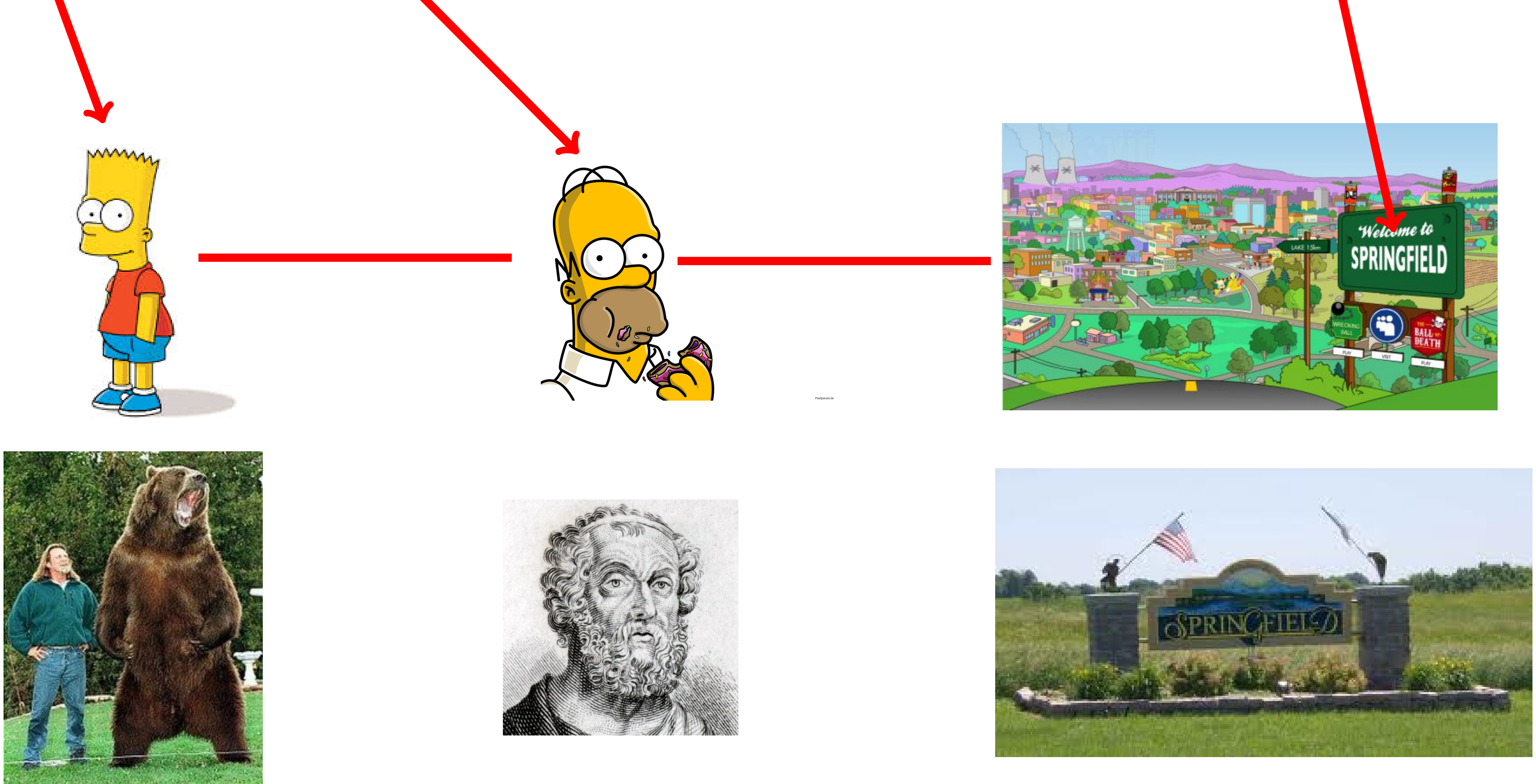
Bart and Homer accidentally launch a rocket into the Springfield church, causing Lisa to leave Christianity.



Def: Coherence Criterion

The Coherence Criterion postulates that entities that are mentioned in one document should be related in the KB.

Bart and Homer accidentally launch a rocket into the Springfield church, causing Lisa to leave Christianity.



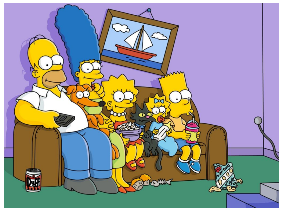
Summary: Disambiguation

We saw 3 indicators for disambiguation:

1. Context

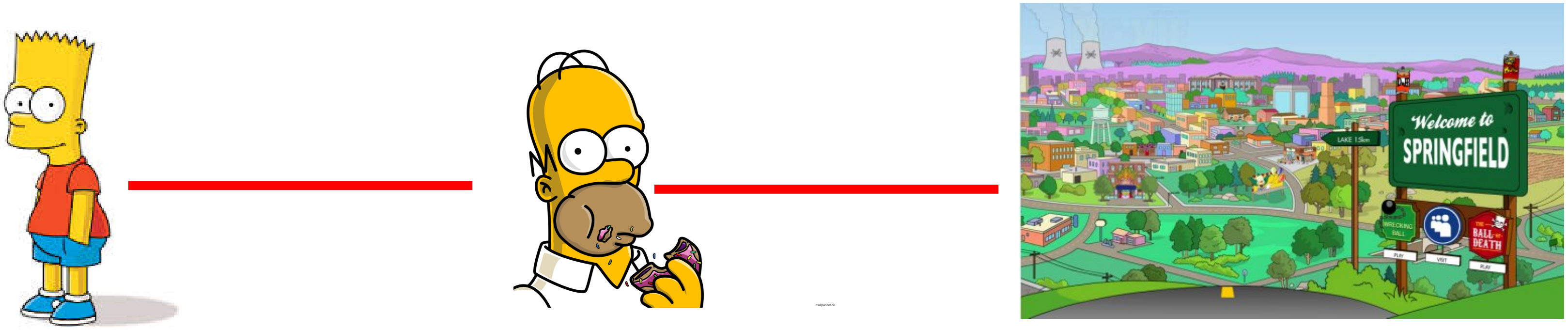
Homer eats a doughnut.

2. Disambiguation prior



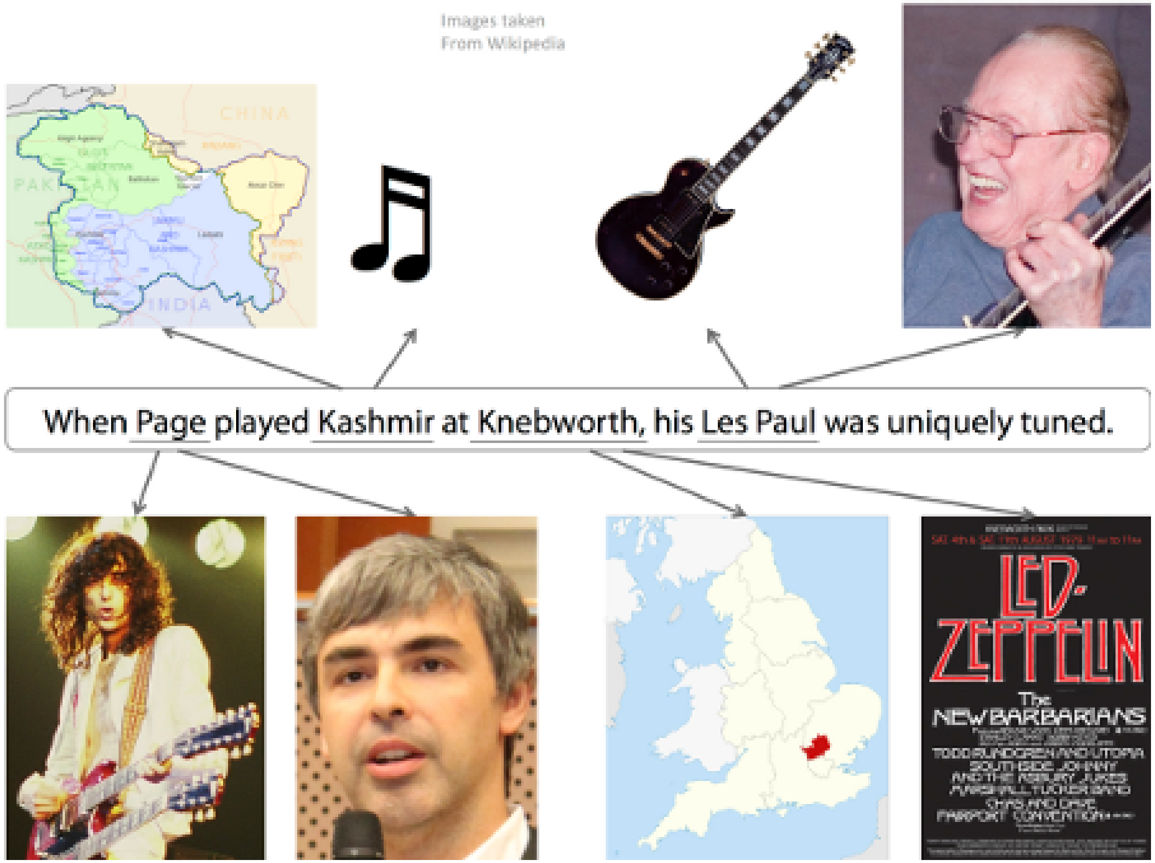
$>$ Simpsons

3. Coherence



Example: Disambiguation by AIDA

AIDA is a system for the disambiguation of entity names, based on YAGO.

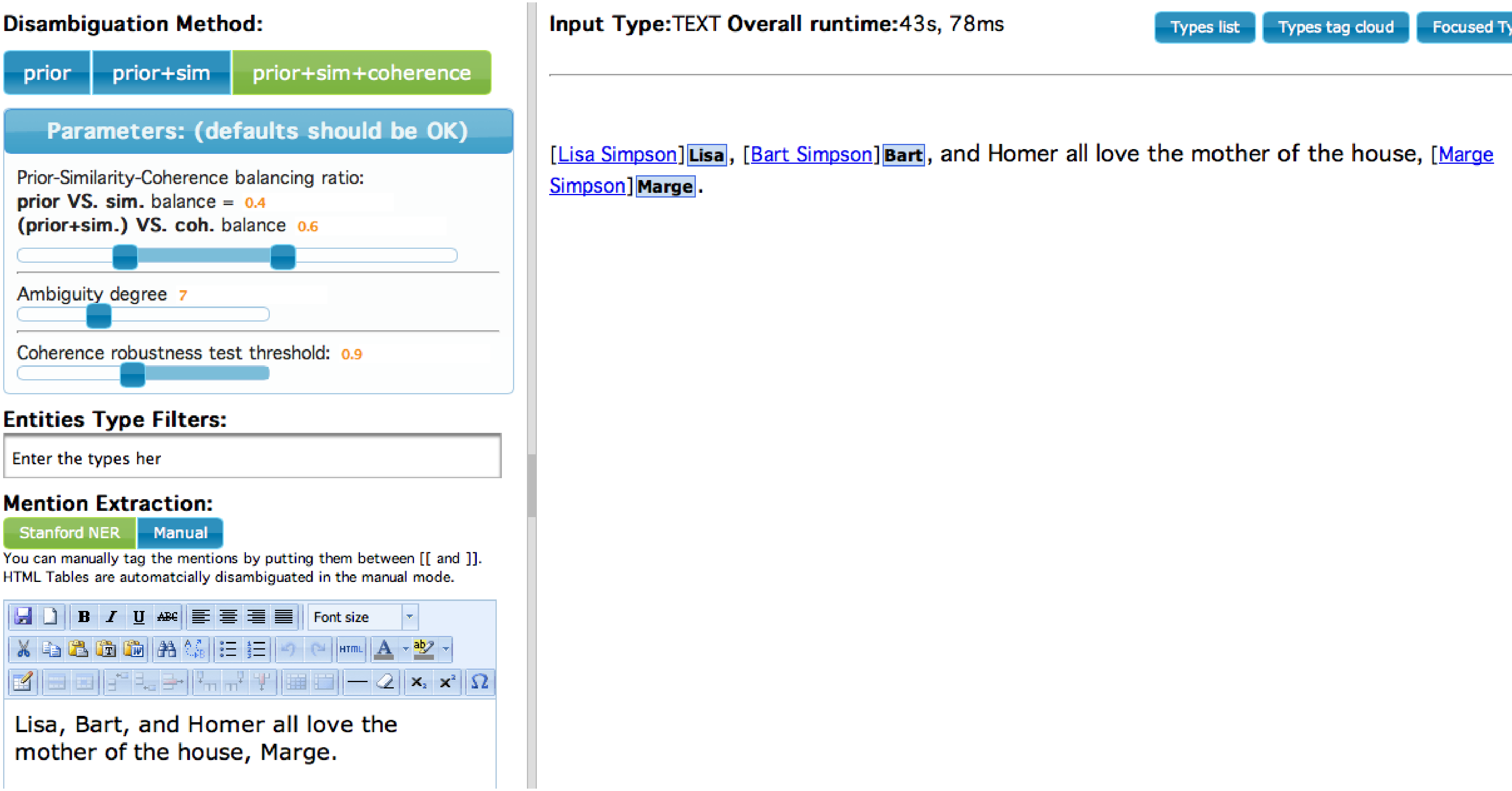


Images taken From Wikipedia

When Page played Kashmir at Knebworth, his Les Paul was uniquely tuned.

Try it out

Example: Disambiguation by AIDA



Disambiguation Method:

Parameters: (defaults should be OK)

Prior-Similarity-Coherence balancing ratio:
prior VS. sim. balance = 0.4
(prior+sim.) VS. coh. balance 0.6

Ambiguity degree 7

Coherence robustness test threshold: 0.9

Entities Type Filters:

Enter the types her

Mention Extraction:

You can manually tag the mentions by putting them between [[and]].
HTML Tables are automatcially disambiguated in the manual mode.

Lisa, Bart, and Homer all love the mother of the house, Marge.

Input Type:TEXT **Overall runtime:**43s, 78ms

[Lisa Simpson]**Lisa**, [Bart Simpson]**Bart**, and Homer all love the mother of the house, [Marge Simpson]**Marge**.

References

AIDA: An Online Tool for Accurate Disambiguation

->instance-extraction