Master Thesis

Named Entity Recognition and Disambiguation with Wikidata on Arbitrary Engilish Text

Yi-Chun Lin 2021/09/22

Problem

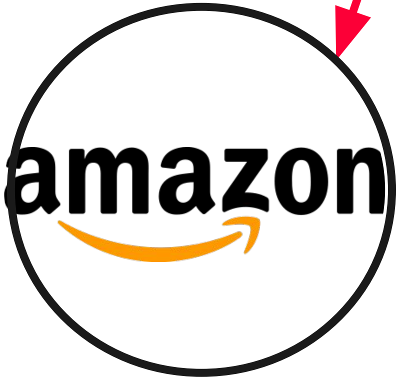
The Tasks

Named Entity Recognition (NER): given a text, tell which words belongs to a named entity

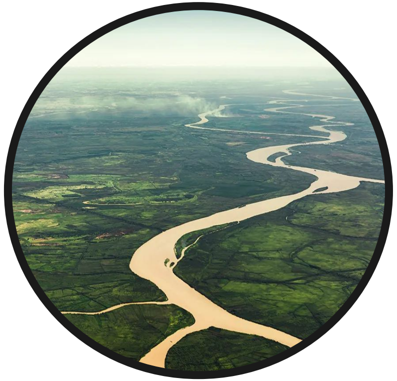
Namd Entity Disambiguation (NED): given the text span of a named entity, link it to its corresponding entry in a knowledge base.

NER + NED Example

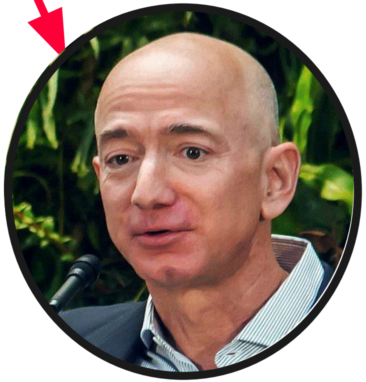
Amazon was founded by Jeff Bezos.



Q3884 (wikidata)



Q3783



Q312556

Problem Definition

input: plain text

Amazon was founded by Jeff Bezos.

output: text span of named entities + corresponding entries in Wikidata

“Amazon”: Q3884,
“Jeff Bezos”: Q312556

Solution

Base Model

WIKIDATA

Entity Index

Candidate List

Query Text

NER

Examine all possible text spans in query, speed up with POS-tag filter

NED

Consider popularity of each candidate and its similarity to query context

Output

Entity Index

key: entity name and synonyms

value: candidate lists
(all entities that have the name or sysnonym)

QID	Name	Synonyms
Q3884	Amazon	Amazon.com
Q3783	Amazon	Amazon River

"Amazon": [Q3884, Q3783]
"Amazon.com": [Q3884]
"Amazon River": [Q3783]

NER - 1/2

Task: locate the text span of named entities in the query text.

Basic approach: part-of-speech (POS) tagging. Determine the grammatical category of each word.

✓	Amazon	was	founed	by	Jeff	Bezos	✗	United	States	of	America
	NNP	*VBD*	*VBN*	*IN*	*NNP*	*NNP*		*NNP*	*NNP*	*IN*	*NNP*

NER - 2/2

Our approach: examine text spans starting with *NNP* or *NN*. Choose the longest match.

Amazon ~~was~~ ~~founded~~ ~~by~~ Jeff Bezos

NNP VBD VBN IN NNP NNP

"Amazon", "Amazon was", … ,
"Amazon was founded by Jeff Bezos",
"Jeff", "Jeff Bezos"

United States ~~of~~ America

NNP NNP IN NNP

"United", "United States",
"United States of",
"United States of America"

NED - 1/3

Task: afer NER, for each text span, choose the most suitable item among the candidates.

Basic approach: choose the most popular candidate.

✓ Obama was the president of the US.
✗ Obama is a city in Japan.

NED - 2/3

Query context plays an important role.

How to measure the similarity between context and each candidate:

- "context": *NN* and *NNP* in the query
- "candidate": words in its name, synonym and description
- "similarity": overlaps between the two

NED - 3/3

Our approach: choose the candidate with the highest score, where score =

popularity_score + **similarity_score**

sitelinks $(0 \sim P_{max})$

overlaps x $\begin{cases} P_{max} / 3 \ * \\ P_{max} / 2 \end{cases}$

* when longer contexts (> 10 words)

Configurable Features

On top of the base model, each feature can be turned on/off to see its effectiveness.

Synonym Expansion	KB Enrichment	FP Reduction
- Family Name - Demonym	- Large DB - Wiki Abstract	- NNP Reduction

Family Name

Problem: Amazon was founded by <u>Bezos</u>.

“Jeff Bezos”: { Q312556 }
“Bezos”: {Q4900382 }

Solution: if an entity is of type “person” and has the property “family name”, add its family name to its synonym. “Bezos”: {Q4900382, Q312556 }

Demonym

Problem: Amazon is an American (adj.) company.

"Demonym" denotes the natives or inhabitants of a particular country, state, city, etc.

Solution: if the entity is a country and has the property "demonym", add its demonym to its synonym. "USA": {Q30}, "American": {Q30}

Large DB

Problem: the condensed version of Wikidata excludes less popular entities, leads to recognition limits.

Solution: try the full version of Wikidata

Wikipedia Abstract

Problem: Wikidata description too short, the algorithm falls back to depend only on popularity.

Solution: represent each entity with the words in corresponding Wikipedia abstract. Adjust weight s.t.

$$\text{similarity score} \propto 1 / \log(\#\text{ words})$$

NNP Reduction

Problem: "Bank Duta" recognized as "Bank" "Duta"

"Bank Duta": an Indonesia bank (not in Wikidata)

"Bank": a film by Charlie Chaplin (in Wikidata)

"Duta" : a family name (in Wikidata)

Solution: remove consequent single-word named entities. Exception: any of them's score > P_{max}.

Quick Recap

WIKIDATA

Entity Index

Candidate List

NER

Examine all possible text spans in query, speed up with POS-tag filter

NED

Consider popularity of each candidate and its similarity to query context

Synonym Expansion

- Family Name
- Demonym

KB Enrichment

- Large DB
- Wiki Abstract

FP Reduction

- NNP Reduction

Evaluation

Datasets

(1) AIDA CoNLL-YAGO

news, manual annotated,
217 words/doc, 20 entities/doc

SOCCER - FRANCE BEAT MEXICO 2-0 IN FRIENDLY . PARIS 1996-08-31 France beat Mexico 2-0 (halftime 0-0) in a friendly soccer international on Saturday . Scorers : Nicolas Ouedec (49th minute) , Youri Djorkaeff (53rd) Attendance : 18,000

(2) ClueWeb12 FACC1

mixed, automatic annotated,
26 words/doc, 1.6 entities/doc,
subset of 50000 docs.

English colonists brought asparagus to North America , but asparagus did not become a commercial crop in the United States until the 19th century .

Metrics

We report Micro F1 and Macro F1 scores.

F1 Score: 2 x **P** x **R** / (**P** + **R**)

Precision: ratio of correctly reported NEs among algorithm output

Recall: ratio of correctly reported NEs among the ground truth

Micro: aggregates data from all documents to compute one score

Macro: one score per document and takes average over all documents

Results of Base Model + Single Feature

configuration	Clueweb Micro F1 / Macro F1	AIDA Micro F1 / Macro F1	memory (GB)
base	39.49 39.98	50.9 50.32	**3.80**
base + family name	39.62 40.48	53.47 52.29	<u>3.89</u>
base + demonym	<u>41.78</u> <u>42.74</u>	**55.65** **56.34**	**3.80**
base + large database	39.86 40.71	51.08 50.68	5.19
base + Wikipedia abstract	38.92 39.29	51.03 50.13	5.86
base + NNP reduction	**47.06** **42.95**	<u>54.26</u> <u>53.22</u>	**3.80**

Effectiveness of Each Feature

configuration	false positive		false negative	
	counts	% change	counts	% change
base	89,305	-	53,194	-
base + family name	90,711	1.57%	51,745	-2.72%
base + demonym	89,066	-0.27%	48,696	**-8.46%**
base + large database	91,659	2.64%	52,056	-2.14%
base + Wikipedia abstract	89,943	0.71%	53,659	0.87%
base + NNP reduction	52,771	**-40.91%**	53,972	1.46%

Comparison to AmbiverseNLU

configuration	Clueweb Micro F1 Macro F1	AIDA Micro F1 Macro F1
enhanced	**49.82** 46.21	61.47 60.98
enhanced + large database	49.61 **46.59**	61.39 61.26
enhanced + Wikipedia abstract	48.78 45.18	60.53 60.17
full	48.55 45.56	62.31 61.50
AmbiverseNLU	44.75 33.58	**68.57** **67.78**

* enhanced = base + family name + demonym + NNP reduction

Comparison to AmbiverseNLU

model	Clueweb			AIDA		
	tp	fp	fn	tp	fp	fn
enhanced	40,176	43,721	37,213	16,340	9,316	11,167
full	40,146	47,833	37,243	17,208	10,522	10,299
AmbiverseNLU	26,083	16,119	48,299	17,136	5,319	10,393