Named-Entity Recognition and Linking with a Wikipedia-based Knowledge Base

Bachelor Thesis Presentation

Niklas Baumert
[niklasbaumert@gmail.com]

2018-11-08

Contents

1 Why and What?—Motivation.

1.1 What is named-entity recognition?

1.2 What is entity linking?

2 How?—Implementation.

2.1 Wikipedia-backed Knowledge Base.

2.2 Entity Recognizer & Linker.

3 Performance Evaluation.

4 Conclusion.

1 Why and What—Motivation.

Built to provide input files for QLever.

- Specialized use-case.
  - General application as an extra feature.
- Providing a *wordsfile* and a *docsfile* as output.
- Wordsfile is input for QLever.
- Support for Wikipedia page titles, Freebase IDs and Wikidata IDs.
  - Convert between those 3 freely afterwards.

The docsfile is a list of all sentences with IDs.

record id	text
1	Anarchism is a political philosophy that advocates [...]
2	These are often described as [...]

The wordsfile contains each word of each sentence and determines entities and their likeliness.

word	Entity?	recordID	score
Anarchism	0	1	1
<Anarchism>	1	1	0.9727
is	0	1	1
a	0	1	1
political	0	1	1
philosophy	0	1	1
<Philosophy>	1	1	0.955

Named-entity recognition (NER) identifies nouns in a given text.

- Finding and categorizing entities in a given text. [1][2]
- (Proper) nouns refer to entities.
- Problem: Words are ambiguous. [1]
  - Same word, different entities within a category—”John F. Kennedy”, the president or his son?
  - Same word, different entities in different categories—”JFK”, person or airport?

Entity linking (EL) links entities to a database.

- Resolve ambiguity by associating mentions to entities in a knowledge base. [3][4][5]
- Determine if “Washington” means “Washington D.C.” or “George Washington” etc.

An example for named-entity recognition and entity linking.

Word	Barack	Obama	served	as	the	President	of	the	United	States	.
Tag	NOUN	NOUN	VERB	ADP	DET	NOUN	ADP	DET	NOUN	NOUN	PUNCT
Entity?	✔		✘	✘	✘	✔					✘
Category	president		—	—	—	official post					—
Link	Barack_Obama		—	—	—	President_of_the_United_States					—

2 How—Implementation.

2.1 Wikipedia-backed Knowledge Base

The knowledge base captures relations between strings and Wikipedia pages.

- Knowledge base creation was part of my Bachelor Project.
- Inspired by *Crosswikis*. [4]
- Extract *normalized strings ⇒ Wikipedia pages* relationships.
- 3 sources of information:
  1. The page title.
  2. Links inside the page.
  3. The infobox.
- Stores ratio how frequent a string refers to a Wikipedia page.

Barack Obama

From Wikipedia, the free encyclopedia
(Redirected from Barrack Obama)

"Barack" and "Obama" redirect here. For other uses, see Barack (disambiguation) and Obama (disambiguation).

Barack Hussein Obama II (/bəˈrɑːk huːˈseɪn oʊˈbɑːmə/ listen);[1] born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate.

Obama was born in 1961 in Honolulu, Hawaii, two years after the territory was admitted to the Union as the 50th state. Raised largely in Hawaii, he also spent one year of his childhood in Washington state and four years in Indonesia. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago. In 1988, he enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. After graduating, he became a civil rights attorney and a professor, teaching constitutional law at the University of Chicago Law School from 1992 to 2004. He represented the 13th district for three terms in the Illinois Senate from 1997 to 2004, when he ran for the U.S. Senate. He received national attention in 2004 with his March primary win, his well-received July Democratic National Convention keynote address, and his landslide November election to the Senate. In 2008, he was nominated for president a year after his campaign began and after a close primary campaign against Hillary Clinton. He was elected over Republican John McCain and was inaugurated on January 20, 2009. Nine months later, he was named the 2009 Nobel Peace Prize laureate, accepting the award with the caveat that he felt there were others "far more deserving of this honor than I."

Barack Obama

44th President of the United States

In office
January 20, 2009 – January 20, 2017

Vice President Joe Biden

Preceded by George W. Bush

Succeeded by Donald Trump

Infobox of Barack Obama.

2.2 Entity Recognizer & Linker “WiNERLi”

Wikipedia Named-Entity Recognition and Linking

The 4 Requirements.

1. Direct Mentions.
2. Partial Mentions.
3. Pronouns—”He, She, It”.
4. Categories—”The <Category>”.

The 4 Requirements—An Example.

Barack Obama(1) is an American politician(1) who served as the 44th President of the United States(1). Obama(2) was born in 1961 in Honolulu, Hawaii(1). He(3) received national attention(1) in 2004 with his March(1) primary win. The president(4) signed many landmark bills(1) into law(1).

Built in Python 3, utilizing spaCy for part-of-speech tagging and sentence detection.

- Uses Python 3 and a bit Cython.
- spaCy does the heavy lifting.
  - Part-of-speech tagging.
  - Sentence detection.
  - But not NER (obviously).
- Works sentence by sentence.

Building sub-sequences and try to fail early to skip forward.

- Build sub-sequences from the current sentence.
- 2 basic rules:
  1. If the first token in the sub-sequence is a punctuation symbol, start a new sub-sequence with the next token.
  2. If the last token in the sub-sequence is an adposition (in, to, during, of, etc.) expand the sub-sequence by the next token.
- Partial data, pronoun and category data is cached.
- “Specificity” = number of words that make up the entity.
- 2 possible entities share words ⇒ use the one with higher specificity.

Direct Mentions.

- Trivial case.
- Check if the last token is a (proper) noun.
  - If true → query the knowledge base for the whole sub-sequence.
    - If the query retrieves a result → expand the sub-sequence.
    - Else → start a new sub-sequence with:
      - The last token, if the length of the sub-sequence is ≥ 2. Because it has to be a noun.
      - The next token, otherwise.
  - If false → start a new sub-sequence with the next token.

Direct Mentions—An example:

- “Water” power → Query={Water} → “Water power” next.
- “Water power” → Query={Hydropower}.
- “Stone power” drill → Query={} → “power” next.
- “power” drill → Query={} → “drill” next.

Partial Mentions.

- Previous direct mention of the entity required.
- Only for people.
  - “The Bank of Scotland(1) is a bank(2) located in Scotland(2).”
- Check tokens against cached data of previous entities.
  - If none exist → query the knowledge base.
- Cache Example:
  - “Barack” → “Barack_Obama”
  - “Obama” → “Barack_Obama”

Pronouns.

- Only working with “He”, “She” and “It”.
- Using a Person⇒Gender mapping.
  - Default to “It” if no mapping exists.
- Check if the last token is a pronoun.
- Use the entity stored for the pronoun.
- Only last mentioned entity is kept in cache for each pronoun.
  - “Her doctorate was in the field of quantum chemistry. It was about...”

Categories.

- Uses the Infobox type to infer categories.
  - ~2600 categories.
- Using a Wikipedia page title⇒Category mapping.
- Checks if “the” is in front of the sub-sequence.
  - If true → look at the cache for the sub-sequence string.

3 Performance Evaluation.

2 datasets—GMB-Walia & hand-crafted Wikipedia.

- GMB-Walia
  - “Annotated Corpus for Named Entity Recognition—Feature Engineered Corpus annotated with IOB and POS tags” by User Walia on kaggle.com.[6]
  - Based on the Groningen Meaning Bank (GMB)[7] from the University of Groningen.
  - Over 45000 sentences.
  - With part-of-speech tags, entities and entity categories.
- Hand-crafted Wikipedia
  - Created by myself.
  - Primarily used to evaluate entity linking.
  - Only 28 sentences, 667 words and 189 entities.

3 Tests—Entity detection, categorization and linking.

- Entity Detection—Identify a string as an entity.
- Entity Categorization—Categorize an entity.
- Entity Linking—Link an entity to a database (here: Wikipedia).
- Measuring Precision, Recall and F1-Score.

WiNERLi improves entity detection rates compared to spaCy.

Dataset	System	Precision	Recall	F1
Wikipedia	WiNERLi	**0.5746**	**0.4074**	**0.4768**
	spaCy	0.5000	0.1111	0.1818
GMB-Walia	WiNERLi	**1.0**	**0.2353**	**0.3810**
	spaCy	**1.0**	0.0883	0.1622

WiNERLi performs worse at entity categorization.

Dataset	System	Precision	Recall	F1
Wikipedia	WiNERLi	**0.5588**	0.1011	0.1712
	spaCy	0.4717	**0.1330**	**0.2075**
GMB-Walia	WiNERLi	**0.5258**	0.3115	0.3912
	spaCy	0.5001	**0.5025**	**0.5013**

The entity linking performance was surprisingly low for the given dataset.

Dataset	System	Precision	Recall	F1
Wikipedia	WiNERLi	0.4184	0.2169	0.2857
	spaCy	N/A		

4 Conclusion.

There is more potential, but it requires changing core assumptions.

- Working on the complete Wikipedia takes time.
  - Better solution to transform Wikimedia Markup into usable text—currently mostly multiple regex.
  - Using an in-memory database could increase the performance.
- Using the Infoboxes isn’t practical.
  - Other source of categorization required.
  - Wikipedia has categories, but they are highly specific, too and need minimization.
- More pronouns should be handled.
- Is “specificity” a good deciding measure?

In summary...

- The knowledge base extracts information from the page title, links in the article and the Infobox.
- The knowledge base can be used for named-entity recognition and entity linking.
- WiNERLi can detect entities as direct and partial mentions and by pronouns and categories.
- WiNERLi is better than spaCy at entity detection.
- WiNERLi is worse than spaCy at entity categorization.
- WiNERLi performs worse than I expected at entity linking.
- Further optimization may be required.

Sources

[1] Jurafsky, D. and Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[2] Zitouni, I. (2014). *Natural Language Processing of Semitic Languages*. Theory and Applications of Natural Language Processing. Springer.

[3] Raiman, J. and Raiman, O. (2018). Deeptype: Multilingual entity linking by neural type system evolution.

[4] Spitkovsky, V. I. and Chang, A. X. (2012). A cross-lingual dictionary for english wikipedia concepts. In *LREC*, pages 3168–3175.

[5] Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics.

[6] kaggle.com/abhinavwalia95/entity-annotated-corpus, Version 4

[7] gmb.let.rug.nl