Natural Language Processing

SoSe 2017

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Named-entity Recognition

Dr. Mariana Neves

June 28th, 2017

Motivation

- Factual information and knowledge are normally expressed by named entities.
  - Who, Whom, Where, When, Which, ...
- It is the core of the information extraction systems.

Applications

- Finding the main information of a company from its reports
  - Founder, Board members, Headquarters, Profits

Siemens

From Wikipedia, the free encyclopedia

For other uses of "Siemens", see Siemens (disambiguation).

Siemens AG (German pronunciation: [ˈziːmɛns]) is a German multinational conglomerate company headquartered in Berlin and Munich. It is the largest engineering company in Europe. The principal divisions of the company are *Industry*, *Energy*, *Healthcare*, and *Infrastructure & Cities*, which represent the main activities of the company. The company is a prominent maker of medical diagnostics equipment and its medical health-care division, which generates about 12 percent of the company's total sales, is its second-most profitable unit, after the industrial automation division.[2]

Siemens and its subsidiaries employ approximately 343,000 people worldwide and reported global revenue of around €71.9 billion in 2014 according to their annual report.

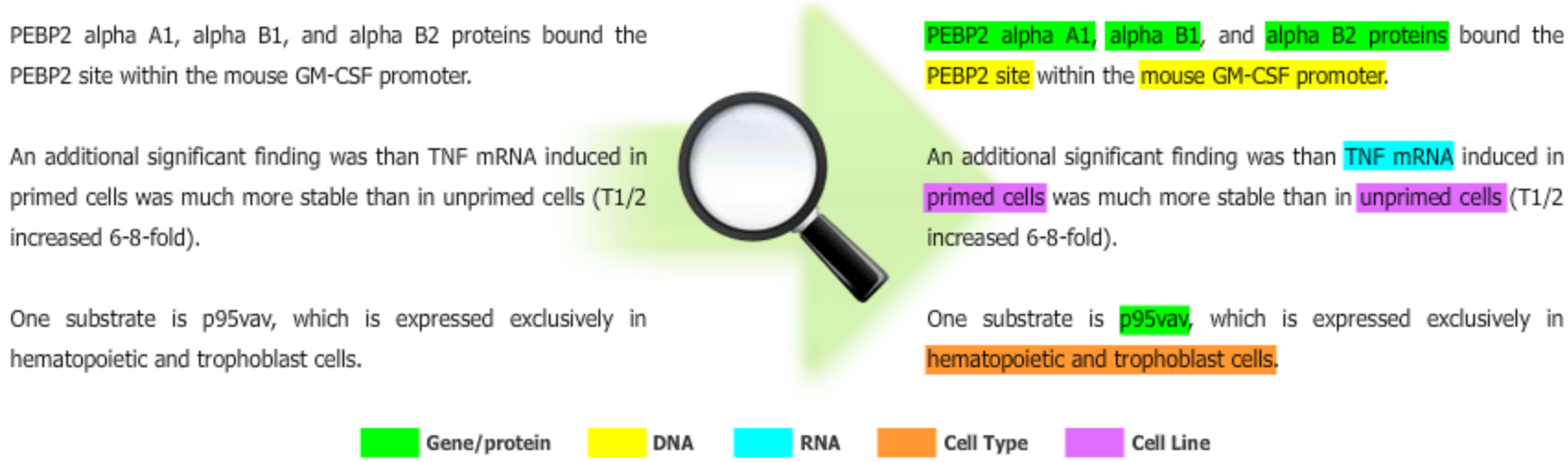
1847 to 1901 [edit]

Siemens & Halske was founded by Werner von Siemens and Johann Georg Halske on 12 October 1847. Based on the telegraph, his invention used a needle to point to the sequence of letters, instead of using Morse code. The company, then called *Telegraphen-Bauanstalt von Siemens & Halske*, opened its first workshop October 12.

(http://en.wikipedia.org/wiki/Siemens)

Applications

- Finding information from biomedical literature
  - Drugs, Genes, Interaction products



(http://bioinformatics.ua.pt/software/gimli/)

Applications

- Finding the target of sentiments
  - Products, Celebrities

An Android User's Review of the iPhone 6: Read this if you are thinking of switching from Android!!!

By Jay on January 28, 2015

Color Name: Gray | Size Name: 16 GB

So I made the switch from using Android to the iPhone back in October, and I've been using the iPhone 6 for the past few months now and can give a detailed review on what it's like to switch over. Before this switch, I've used the Samsung Galaxy S2 (first smartphone ever!) and also the Nexus 4. Since I'm a tech enthusiast, I'm well versed and have played around with many other Android devices, including all the big names, Galaxy S5, HTC One M8 and M7, One Plus One, and so on. Here are my thoughts:

Things that the iPhone does really well (both hardware and software-wise):
1. Camera. The behind the scene software for digitally capturing an image is definitely the strongest sell for the iPhone. Other than the S5 and Note 4, no smartphone really comes close to having the same kind of image quality (no matter the megapixels) compared to the iPhone. This was one of the reasons for me to switch over since I've started to dabble with photography and wanted a really good camera in my smartphone. (Side note, if you read a lot of tech blogs, there is a notion that in the near future our smartphones won't accurately describe our devices anymore since making a phone call is probably one of the least commonly used features on a smartphone when you look at any average user. Cameras, social media, emails all take a higher usage rate than making a call... really interesting, but anyway, back to the review).

2. Reliability. There have been maybe 2 or 3 times when my phone crashed and would have to be restarted, mostly due to playing some game that was not written very well for the iOS devices. Read more ›

(http://www.amazon.com/Apple-iPhone-Space-Gray-Unlocked/dp/B00NQGP42Y/ref=sr_1_1?s=wireless&ie=UTF8&qid=1431337473&sr=1-1&keywords=iphone)

Named Entity Recognition (NER)

- Finding named entities in a text
- Classifying them to the corresponding classes
- Assigning a unique identifier from a database

- „Steven Paul Jobs, co-founder of Apple Inc, was born in California.”
- „Steven Paul Jobs, co-founder of Apple Inc, was born in California.”
- „Steven Paul Jobs [PER], co-founder of Apple Inc [ORG], was born in California [LOC].”
- „Steven Paul Jobs [Steve_Jobs], co-founder of Apple Inc [Apple_Inc.], was born in California [California].”

Named Entity Classes

- Person
  - Person names
- Organization
  - Companies, Government, Organizations, Committees, ..
- Location
  - Cities, Countries, Rivers, ..
- Date and time expression
- Measure
  - Percent, Money, Weight, ...
- Book, journal title
- Movie title
- Gene, disease, drug name

Named Entity Classes (IO)

Steven	PER
Paul	PER
Jobs	PER
,	O
co-founder	O
of	O
Apple	ORG
Inc	ORG
,	O
was	O
born	O
in	O
California	LOC
.	O

Named Entity Classes (BIO/IOB)

Steven	B-PER
Paul	I-PER
Jobs	I-PER
,	O
co-founder	O
of	O
Apple	B-ORG
Inc	I-ORG
,	O
was	O
born	O
in	O
California	B-LOC
.	O

Named Entity Classes (BIEWO)

Steven	B-PER
Paul	I-PER
Jobs	E-PER
,	O
co-founder	O
of	O
Apple	B-ORG
Inc	E-ORG
,	O
was	O
born	O
in	O
California	W-LOC
.	O

NER Ambiguity (IO vs. IOB encoding)

John	PER
shows	O
Mary	PER
Hermann	PER
Hesse	PER
's	O
book	O
.	O

John	B-PER
shows	O
Mary	B-PER
Hermann	B-PER
Hesse	I-PER
's	O
book	O
.	O

NER Ambiguity

- Ambiguity between named entities and common words
  - May: month, verb, surname
  - Genes: VIP, hedgehog, deafness, wasp, was, if
- Ambiguity between named entity types
  - Washington (Location or Person)

NER task

- Similar to a classification task
  - Feature selection
  - Algorithms

Feature Selection

- Features
  - Word:
    - Germany: Germany
  - POS tag:
    - Washington: NNP
  - Capitalization:
    - Stefan: [CAP]

Feature Selection

- Features
  - Punctuation:
    - St.: [PUNC]
  - Lowercased word:
    - Book: book
  - Suffixes:
    - Spanish: -ish
  - Word shapes:
    - 1920-2008: dddd-dddd
    - ABC-123: AAA-ddd

NER

- List lookup
  - Extensive list of names are available via various resources
  - Gazetteer: a large list of place names
  - Biomedical: database of genes, proteins, drugs names
  - Usually good precision (depending on the domain), but low recall (there are many variations)

Sequence Modeling

- Many of the NLP techniques should deal with data represented as sequence of items
  - Characters, Words, Phrases, Lines, ...

- I[PRP] saw[VBP] the[DT] man[NN] on[IN] the[DT] roof[NN] .

- Steven[PER] Paul[PER] Jobs[PER] ,[O] co-founder[O] of[O] Apple[ORG] Inc[ORG] ,[O] was[O] born[O] in[O] California[LOC].

Sequence Modeling

- Making a decision based on the
  - Current Observation
    - Word (W0)
    - Prefix
    - Suffix
    - Lowercased word
    - Capitalization
    - Word shape

Sequence Modeling

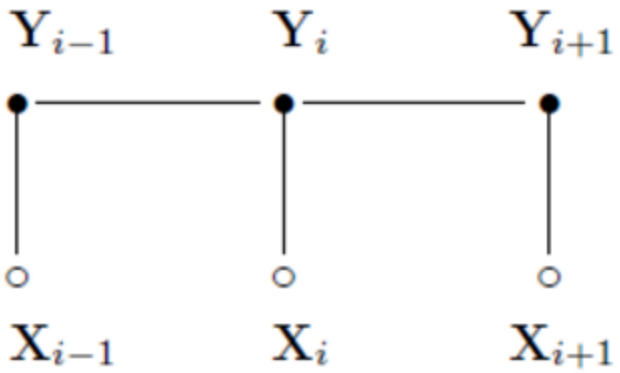
- Making a decision based on the
  - Surrounding observations
    - W+1
    - W−1
  - Previous decisions
    - T−1
    - T−2

Algorithms

- Hidden-Markov Models (HMM)
  - Same used for POS tagging
- Conditional Random Fields (CRF)
- Machine learning algorithms

Conditional Random Fields (CRF)

- Discriminative undirected probablistic graphical model
- Linear chain CRF
  - Sequential classification model
  - Predicts sequence of labels for sequences of input samples



(http://curtis.ml.cmu.edu/w/courses/index.php/Lafferty_2001_Conditional_Random_Fields)

Neural networks

- CRF-LSTM architecture

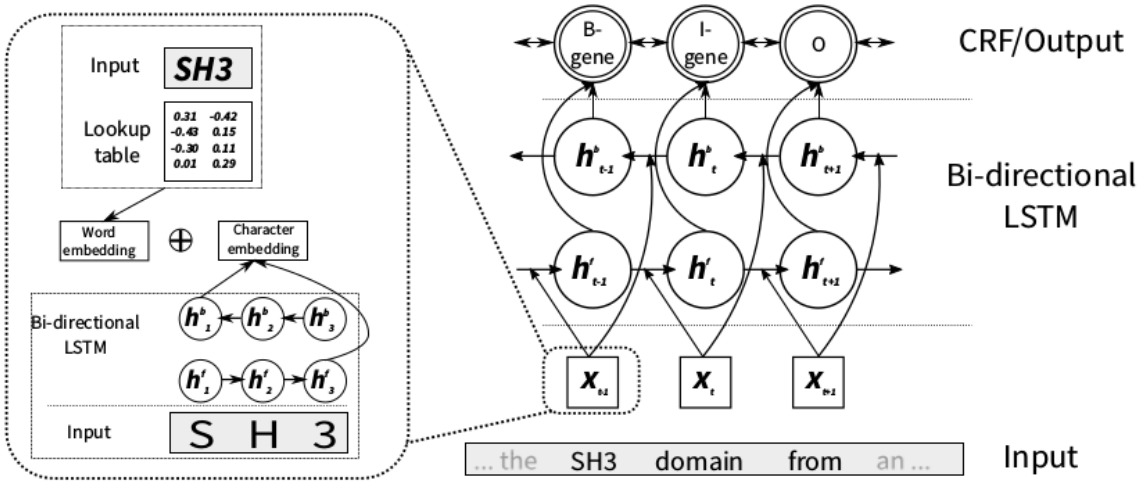


Fig. 1: CRF-LSTM architecture. For instance, for the word w_{t-1} ="SH3" from the input sentence S, the character-based representation is computed by applying a bi-directional LSTM onto the sequence of its characters 'S', 'H', '3'. The resulting embedding is concatenated with the corresponding word embedding, trained on a huge corpus. This word representation is then processed by another bi-directional LSTM and finally by a CRF layer. The output is the most probable tag sequence, as estimated by the CRF.

(Habibi M, Weber L, Neves M, Wiegandt D L and Leser U.
Deep Learning with Word Embeddings improves Biomedical Named Entity Recognition, ISMB/ECCB 2017, accepted, Prague, Czech Republic.)

Evaluation

- Precision, recall and F-measure metrics based on
  - True positive
    - Correctly identified mentions
  - False positive
    - Incorrectly identified mentions
  - False negative
    - Correct mentions which were missed

Summary

- Named-entity recognition
- Various entity classes and encoding
- Two main approches
  - Dictionary-based
  - Sequential labeling (machine learning)

Further Reading

- Speech and Language Processing book
  - Chapter 22.1: NER

- CRF: Lafferty, McCallum and Pereira 2001:
  - http://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers