Sequence Segmentation Models

Yoav Goldberg

Bar Ilan University

Last time

- Discriminative Classification:

$$f(x) = \arg\max_{y} score(x, y)$$

Last time

- Discriminative Classification:

$$f(x) = \arg\max_y score(x, y)$$

  - Linear classifier:

$$f(x) = \arg\max_y score(x, y) = \arg\max_y \color{red}{w \cdot \phi(x, y)}$$

Last time

- Discriminative Classification:

$$f(x) = \arg\max_y score(x, y)$$

  - Linear classifier:

$$f(x) = \arg\max_y score(x, y) = \arg\max_y w \cdot \phi(x, y)$$

  - Log-linear logistic-regression (MaxEnt) classifier:

$$f(x) = \arg\max_y score(x, y) = \arg\max_y p(y|x) =$$

Last time

- Discriminative Classification:

$$f(x) = \arg\max_{y} score(x, y)$$

  - Linear classifier:

$$f(x) = \arg\max_{y} score(x, y) = \arg\max_{y} w \cdot \phi(x, y)$$

  - Log-linear logistic-regression (MaxEnt) classifier:

$$f(x) = \arg\max_{y} score(x, y) = \arg\max_{y} p(y|x) = \arg\max_{y} \frac{e^{w \cdot \phi(x,y)}}{\sum_{y'} e^{w \cdot \phi(x,y')}}$$

Side Note – Linear Models

Reminder:

$$f(x) = w \cdot \phi(x)$$

Side Note – Linear Models

Reminder:

$$f(x) = w \cdot \phi(x) = w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_m\phi_m(x)$$

Side Note – Linear Models

Reminder:

$$f(x) = w \cdot \phi(x) = w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_m\phi_m(x)$$

What cannot be captured?

Side Note – Linear Models

Reminder:

$$f(x) = w \cdot \phi(x) = w_1\phi_1(x) + w_2\phi_2(x) + \cdots + w_m\phi_m(x)$$

What cannot be captured?

$$x = \text{person}$$

$\phi_1 \ldots \phi_{10}$ look at height (1 to 10, 1 very low, 10 very high)
$\phi_{11} \ldots \phi_{20}$ looks at weight (1 to 10, 1 very low, 10 very high)

Need to classify "has eating disorder"

Back to NLP

Last time: the sequence tagging problem

Input

Holly came from Miami , F.L.A ,
hitch-hiked her way across the USA

Output

Holly/NNP came/VBD from/IN Miami/NNP ,/, F.L.A/NNP ,/,
hitch-hiked/VBD her/PRP way/NN across/IN the/DT USA/NNP

Assign a **tag** from a given **tagset** to each **word** in a sentence.

Part-of-speech Tagging: Discussion

- Part-of-speech tagging is a “solved” problem.
- It is robust, well understood, and performs well.
- There are many off-the-shelf, downloadable tools:
  - TnT
  - Stanford Tagger
  - SVMTool
  - ClearNLP
  - . . . many others

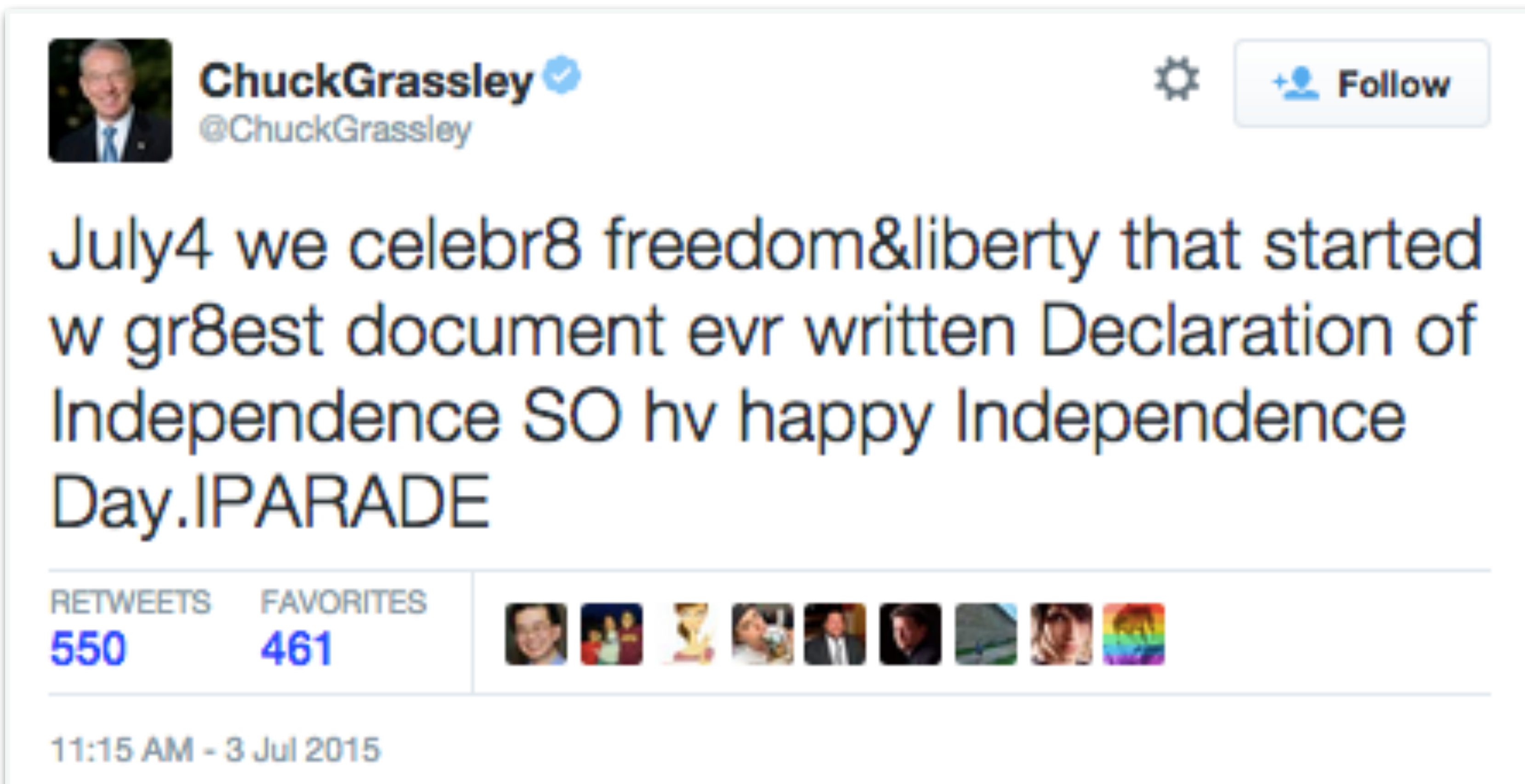
Part-of-speech Tagging: Discussion

- Part-of-speech tagging is a “solved” problem.
- It is robust, well understood, and performs well.
- There are many off-the-shelf, downloadable tools:
  - TnT
  - Stanford Tagger
  - SVMTool
  - ClearNLP
  - . . . many others
- Not all is good, though.
  - Drop in accuracy when moving to a new domain.
    - web, reviews, email, fiction, medical, old text, . . .

Part-of-speech Tagging: Discussion

- Part-of-speech tagging is a “solved” problem.
- It is robust, well understood, and performs well.
- There are many off-the-shelf, downloadable tools:
  - TnT
  - Stanford Tagger
  - SVMTool
  - ClearNLP
  - . . . many others
- Not all is good, though.
  - Drop in accuracy when moving to a new domain.
    - web, reviews, email, fiction, medical, old text, . . .
  - Languages other than English.
    - What if there is not much training data available?
    - Inflections (many word forms $\rightarrow$ OOV, but also more hints)
    - Words may not be white-space delimited: בבית, ומספר.

Bad Language?



Noah Smith
@nlpnoah Pittsburgh, PA
http://www.cs.cmu.edu/~nasmith

omg, first tweet evar! I'm in the green room at #SXSW getting ready for my panel, #textworld

13 Mar via web

Favorite Retweet Reply

Also: at-mentions, URLs, emoticons, symbols, typos, etc.

Carnegie Mellon

Penn Treebank tokenization is unsuitable for Twitter:

```
@user1 OMG ur from PA ? i am too (: where abouts ?
```

you're

I'm going to

```
@user2 ima get me a flip phone for real
```

Penn Treebank tokenization is unsuitable for Twitter:

nominal+verbal

`@user1 OMG ur from PA ? i am too (: where abouts ?`

you're

I'm going to

`@user2 ima get me a flip phone for real`

nominal+verbal

Solution:

Don't try to tokenize these

Instead, introduce **compound tags**

Hashtags

Twitter hashtags are sometimes used as ordinary words (35% of the time) and other times as topic markers

proper noun

```
Innovative , but traditional , too ! Another
fun one to watch on the #iPad !
http://bit.ly/ @user1 #utcd2 #utpol #tcot
```

hashtag

We only use "hashtag" for topic markers

Twitter Discourse Marker

Retweet construction:

```
RT @user1 : I never bought candy bars from
those kids on my doorstep so I guess they're
all in gangs now .
```

Twitter Discourse Marker

Retweet construction:

RT @user1 : I never bought candy bars from those kids on my doorstep so I guess they're all in gangs now .

Twitter discourse marker

RT @user2 : LMBO ! This man filed an EMERGENCY Motion for Continuance on account of the Rangers game tonight ! ⟨⟨ Wow lmao

ARK lti Carnegie Mellon

- Twitter-specific tags:

  hashtag

  at-mention

  URL / email address

  emoticon

  Twitter discourse marker

  other (multi-word abbreviations, symbols, garbage)

Today: Sequence Segmentation

Pierre Vinken , 61 years old , will join the board as a nonexecutive director , Nov. 21, 1987

$\Downarrow$ chunking

Today: Sequence Segmentation

Pierre Vinken , 61 years old , will join the board as a nonexecutive director , Nov. 21, 1987

$\Downarrow$ chunking

$\overbrace{\text{Pierre Vinken}}^{\text{NP}}$, $\overbrace{\text{61 years old}}^{\text{ADJP}}$, $\overbrace{\text{will join}}^{\text{VP}}$ $\overbrace{\text{the board}}^{\text{NP}}$ $\overbrace{\text{as}}^{\text{PP}}$
$\overbrace{\text{a nonexecutive director}}^{\text{NP}}$, $\overbrace{\text{Nov. 21, 1987}}^{\text{NP}}$.

Today: Sequence Segmentation

Pierre Vinken , 61 years old , will join the board as a nonexecutive director , Nov. 21, 1987

$\Downarrow$ chunking

$\overbrace{\text{Pierre Vinken}}^{\text{NP}}$, $\overbrace{\text{61 years old}}^{\text{ADJP}}$, $\overbrace{\text{will join}}^{\text{VP}}$ $\overbrace{\text{the board}}^{\text{NP}}$ $\overbrace{\text{as}}^{\text{PP}}$
$\overbrace{\text{a nonexecutive director}}^{\text{NP}}$, $\overbrace{\text{Nov. 21, 1987}}^{\text{NP}}$.

alternative notation

[NP Pierre Vinken **]** , **[ADJP** 61 years old **]** , **[VP** will join **]** **[NP** the board **]** **[PP** as **]** **[NP** a nonexecutive director **]** , **[NP** Nov. 21, 1987 **]**.

Today: Sequence Segmentation

Pierre Vinken , 61 years old , will join the board as a nonexecutive director , Nov. 21, 1987

$\Downarrow$ chunking

$\overbrace{\text{Pierre Vinken}}^{\text{NP}}$, $\overbrace{\text{61 years old}}^{\text{ADJP}}$, $\overbrace{\text{will join}}^{\text{VP}}$ $\overbrace{\text{the board}}^{\text{NP}}$ $\overbrace{\text{as}}^{\text{PP}}$
$\overbrace{\text{a nonexecutive director}}^{\text{NP}}$, $\overbrace{\text{Nov. 21, 1987}}^{\text{NP}}$.

Input: Sequence. **Output**: Typed, non-overlapping spans

Today: Sequence Segmentation

This is a very useful and general task.

Sequence Segmentation

Chunking

Chunking

Pierre Vinken , 61 years old , will join the board as a nonexecutive director , Nov. 21, 1987

$$\Downarrow$$

$$\overbrace{\text{Pierre Vinken}}^{\text{NP}}\text{ , }\overbrace{\text{61 years old}}^{\text{ADJP}}\text{ , }\overbrace{\text{will join}}^{\text{VP}}\ \overbrace{\text{the board}}^{\text{NP}}\ \overbrace{\text{as}}^{\text{PP}}$$

$$\overbrace{\text{a nonexecutive director}}^{\text{NP}}\text{ , }\overbrace{\text{Nov. 21, 1987}}^{\text{NP}}.$$

NP Chunking

Pierre Vinken , 61 years old , will join the board as a nonexecutive director , Nov. 21, 1987

$\Downarrow$

$\overbrace{\text{Pierre Vinken}}^{\text{NP}}$, 61 years old , will join $\overbrace{\text{the board}}^{\text{NP}}$ as
$\overbrace{\text{a nonexecutive director}}^{\text{NP}}$, $\overbrace{\text{Nov. 21, 1987}}^{\text{NP}}$.

Sequence Segmentation

Named Entities

Named Entity Recognition (NER)

Paris Whitney Hilton (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Identify things

Person / **Loc**ation / **Org**anization / **Time** / **Other**.

Sequence Segmentation

Named Entities

Named Entity Recognition (NER)

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Named Entity Recognition (NER)

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Q

What's the difference between NP-chunking and NER? When would you use each?

Bibliographies

Sequence Segmentation

Information Extraction

Aho, Alfred V., and Jeffrey D. Ullman (1972) *the theory of parsing, translation and compiling*, in two volumes, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Knuth, D.E. (1965) "On the translation of language from left to right," *Information and Control* 8.6, 607-609

Sequence Segmentation

Information Extraction

$\overbrace{\text{Aho, Alfred V.}}^{\text{author}}$, and $\overbrace{\text{Jeffrey D. Ullman}}^{\text{author}}$ ($\overbrace{\text{1972}}^{\text{year}}$)
$\overbrace{\textit{the theory of parsing, translation and compiling}}^{\text{title}}$, in two volumes,
$\overbrace{\text{Prentice-Hall, Inc.}}^{\text{publisher}}$, Englewood Cliffs, NJ.

$\overbrace{\text{Knuth, D.E.}}^{\text{author}}$ ($\overbrace{\text{1965}}^{\text{year}}$) “$\overbrace{\text{On the translation of language from left to right,}}^{\text{title}}$”
$\overbrace{\textit{Information and Control}}^{\text{journal}}$ $\overbrace{\text{8.6}}^{\text{volume}}$, $\overbrace{\text{607-609}}^{\text{pages}}$

Sequence Segmentation

Information Extraction

Seminar Announcements

We will begin a new topic on Tuesday. Andzy Zucker will speak on interactions between logic and topological dynamics.

Mathematical logic seminar – November 12 2013
Time: 12:00 – 13:20

Room: Wean Hall 7201

Department of Mathematical Sciences
CMU

Title: Interactions between logic and topological dynamics

The CMU mathematical logic seminar will meet Tue 5 Nov 12:00–1:30 in Wean Hall 7201.

Alexei Kolesnikov will conclude his series of talks, and speak about Generalized Martin's axiom and disjoint amalgamation.

Note to newcomers: Seminar follows a "brown bag lunch" format and for
the first ten minutes or so we chat and have lunch. The actual talk usually runs 12:10-1:20 or so.

We will begin a new topic on $\overbrace{\text{Tuesday}}^{\text{date}}$. $\overbrace{\text{Andzy Zucker}}^{\text{speaker}}$ will speak on $\overbrace{\text{interactions between logic and topological dynamics}}^{\text{topic}}$.

$\overbrace{\text{Mathematical logic seminar}}^{\text{seminar}}$ – $\overbrace{\text{November 12 2013}}^{\text{date}}$

Time: $\overbrace{\text{12:00 – 13:20}}^{\text{time}}$

Room: $\overbrace{\text{Wean Hall 7201}}^{\text{location}}$

Department of Mathematical Sciences
CMU

Title: $\overbrace{\text{Interactions between logic and topological dynamics}}^{\text{topic}}$

$\overbrace{\text{The CMU mathematical logic seminar}}^{\text{seminar}}$ will meet $\overbrace{\text{Tue 5 Nov}}^{\text{date}}$ $\overbrace{\text{12:00–1:30}}^{\text{time}}$ in $\overbrace{\text{Wean Hall 7201}}^{\text{location}}$.

$\overbrace{\text{Alexei Kolesnikov}}^{\text{speaker}}$ will conclude his series of talks, and talk about $\overbrace{\text{Generalized Martin's axiom and disjoint amalgamation}}^{\text{topic}}$.

Note to newcomers: Seminar follows a “brown bag lunch” format and for
the first ten minutes or so we chat and have lunch. The actual talk usually runs 12:10-1:20 or so.

Gene Mention Identification

Beta-endorphin , ACTH and cortisol secretion were measured in twelve healthy adult males after nasal spray administration 200 IU salmon calcitonin .

Sequence Segmentation

Information Extraction

Gene Mention Identification

Beta-endorphin , ACTH and cortisol secretion were measured in twelve healthy adult males after nasal spray administration 200 IU salmon calcitonin .

$$\Downarrow$$

$\overbrace{\text{Beta-endorphin}}$, $\overbrace{\text{ACTH}}$ and cortisol secretion were measured in twelve healthy adult males after nasal spray administration 200 IU $\overbrace{\text{salmon calcitonin}}$.

Bioinformatics

A gene encding a putative human RNA helicase, p54, has been cloned and mapped to the band q23.3 of chromosome 11. The predicted amino acid sequence shares a striking homology (75% identical) with the female germline-specific RNA helicase ME31B gene of Drosophila. Unlike ME31B, however, the new gene expresses an abundant transcript in a large number of adult tissues and its 5‘ non-coding region was found split in a t(11;14)(q23.3;q32.3) cell line from a diffuse large B-cell lymphoma.

P54; RCK; HLR2	P54; NMT55; NRB54	P54; FKBP51; PPlase	S4; dRpt2; p54;p56
Species: H. sapiens	Species: H. sapiens	Species: H. sapiens	Species: D. melanogaster
Chromosome 11q23.3	Chromosome Xq13.1	Chromosome 6p21.3-2	Chromosome 3R;95C13
GO: RNA helicase	GO: RNA splicing	GO: isomerase activity	GO: Proteolysis

Medical – clinical patient notes

She had a liver function test and amylase and lipase postoperatively and she had a digoxin level of 1.0 on 06/04/05 .
The patient had a CBC on admission of 14.1 with a hematocrit of 33.8 .
Her CBC remained stable on 06/05/05 .
She had a white blood cell of 7.7 , hematocrit of 30.6 .
The patient had a MRSA nasal culture obtained on 06/03/05 , which revealed rare staphylococcus aureus .
The patient had a chest x-ray on admission , which was clear .
No pleural effusion or pneumothorax .

source:
`https://www.i2b2.org/NLP/Relations/Documentation.php`

She had a liver function test and amylase and lipase postoperatively and she had a digoxin level of 1.0 on 06/04/05 .
The patient had a CBC on admission of 14.1 with a hematocrit of 33.8 .
Her CBC remained stable on 06/05/05 .
She had a white blood cell of 7.7 , hematocrit of 30.6 .
The patient had a MRSA nasal culture obtained on 06/03/05 , which revealed rare staphylococcus aureus .
The patient had a chest x-ray on admission , which was clear .
No pleural effusion or pneumothorax .

What test were performed?
What problems were looked for?

source:
`https://www.i2b2.org/NLP/Relations/Documentation.php`

She had $\overbrace{\text{a liver function test}}^{\text{test}}$ and $\overbrace{\text{amylase}}^{\text{test}}$ and $\overbrace{\text{lipase}}^{\text{test}}$

postoperatively and she had $\overbrace{\text{a digoxin level}}^{\text{test}}$ of 1.0 on 06/04/05 .

The patient had $\overbrace{\text{a CBC}}^{\text{test}}$ on admission of 14.1 with a $\overbrace{\text{hematocrit}}^{\text{test}}$ of 33.8 .

$\overbrace{\text{Her CBC}}^{\text{test}}$ remained stable on 06/05/05 .

She had $\overbrace{\text{a white blood cell}}^{\text{test}}$ of 7.7 , $\overbrace{\text{hematocrit}}^{\text{test}}$ of 30.6 .

The patient had a $\overbrace{\text{MRSA nasal culture}}^{\text{test}}$ obtained on 06/03/05 ,

which revealed $\overbrace{\text{rare staphylococcus aureus}}^{\text{problem}}$.

The patient had $\overbrace{\text{a chest x-ray}}^{\text{test}}$ on admission , which was clear .

No $\overbrace{\text{pleural effusion}}^{\text{problem}}$ or $\overbrace{\text{pneumothorax}}^{\text{problem}}$.

Sequence Segmentation

Information Extraction

Many other domains

- Housing Adds
- Restaurant reviews
- Camera reviews
- Police reports
- Politics / news
- ...

Evaluation

Sequence Segmentation – Evaluation

- Assume we have a system. How do we know how well we did?
- In tagging, calculate the percentage of correct words
- Why is segmentation different?

Sequence Segmentation – Evaluation

- Assume we have a system. How do we know how well we did?
- In tagging, calculate the percentage of correct words
- Why is segmentation different?

Metric 1

- Percentage of correct sentences.

Sequence Segmentation – Evaluation

- Assume we have a system. How do we know how well we did?
- In tagging, calculate the percentage of correct words
- Why is segmentation different?

Metric 1

- Percentage of correct sentences.
- Problem: too harsh. We need a segment-level metric.

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}}$$

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}}$$

Gold:

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman .

Predicted:

$\overbrace{\text{Paris}}^{\text{LOC}}$ $\overbrace{\text{Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an $\overbrace{\text{American television}}^{\text{ORG}}$ personality and businesswoman .

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}} = \frac{1}{4}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}}$$

Gold:

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman .

Predicted:

$\overbrace{\text{Paris}}^{\text{LOC}}$ $\overbrace{\text{Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an $\overbrace{\text{American television}}^{\text{ORG}}$ personality and businesswoman .

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}} = \frac{1}{4}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}} = \frac{1}{2}$$

Gold:

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman .

Predicted:

$\overbrace{\text{Paris}}^{\text{LOC}}$ $\overbrace{\text{Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an $\overbrace{\text{American television}}^{\text{ORG}}$ personality and businesswoman .

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}}$$

Further Considerations

- What do we consider "correct"?
- ? Off-by-one errors…
- ? Type-errors…

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}}$$

We have two numbers. What do we optimize?

- In general,
  - higher precision $\rightarrow$ lower recall
  - higher recall $\rightarrow$ lower precision

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}}$$

We have two numbers. What do we optimize?

- In general,
  - higher precision $\rightarrow$ lower recall
  - higher recall $\rightarrow$ lower precision
- Need to balance them.

Sequence Segmentation – Evaluation

Solution

Precision

$$\frac{\text{num correct predictions}}{\text{num all predictions}}$$

Recall

$$\frac{\text{num correct predictions}}{\text{num of gold segments}}$$

We have two numbers. What do we optimize?

- In general,
  - higher precision $\rightarrow$ lower recall
  - higher recall $\rightarrow$ lower precision
- Need to balance them.
- F_1

$$F_1 = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Sequence Segmentation – Evaluation

$$F_1 = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

F_1 gives equal weight to precision and recall.
We may prefer one over the other.

Sequence Segmentation – Evaluation

$$F_1 = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

F_1 gives equal weight to precision and recall.
We may prefer one over the other.

In general,

$$F_\beta = (1 + \beta^2)\frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}}$$

Methods

Sequence Segmentation – Methods

Concrete Example – NP Chunking

Data visualization tools are often used to communicate complex information in a more intuitive way . By representing high dimensional data (e.g. objects with 4 or more variables) by two-dimensional points , in such a way that similar objects are represented by nearby points and dissimilar objects are represented by distant points , scientists can gain intuition by examining the presence of structure and clustering in plots of the data .

Sequence Segmentation – Methods

Concrete Example – NP Chunking

Data visualization tools are often used to communicate complex information in a more intuitive way . By representing high dimensional data (e.g. objects with 4 or more variables) by two-dimensional points , in such a way that similar objects are represented by nearby points and dissimilar objects are represented by distant points , scientists can gain intuition by examining the presence of structure and clustering in plots of the data .

Sequence Segmentation – Methods

Concrete Example – NP Chunking

Data visualization tools are often used to communicate complex information in a more intuitive way . By representing high dimensional data (e.g. objects with 4 or more variables) by two-dimensional points , in such a way that similar objects are represented by nearby points and dissimilar objects are represented by distant points , scientists can gain intuition by examining the presence of structure and clustering in plots of the data .

Look at the POS-tags

NN NN NNS VBP RB VBN TO VB JJ NN IN DT JJR JJ NN . IN VBN JJ JJ NN (FW) NNS IN CD CC JJR NNS) IN JJ NNS , in PDT DT NN IN JJ NNS VBP VBD IN JJ NNS CC JJ NNS VBP VBD IN JJ NNS , NNS MD VB NN IN VBG DT NN IN NN JJ NN IN NNS IN DT NN .

Sequence Segmentation – Methods

Concrete Example – NP Chunking

Data visualization tools are often used to communicate complex information in a more intuitive way . By representing high dimensional data (e.g. objects with 4 or more variables) by two-dimensional points , in such a way that similar objects are represented by nearby points and dissimilar objects are represented by distant points , scientists can gain intuition by examining the presence of structure and clustering in plots of the data .

Look at the POS-tags

NN NN NNS VBP RB VBN TO VB JJ NN IN DT JJR JJ NN . IN VBN JJ JJ NN (FW) NNS IN CD CC JJR NNS) IN JJ NNS , in PDT DT NN IN JJ NNS VBP VBD IN JJ NNS CC JJ NNS VBP VBD IN JJ NNS , NNS MD VB NN IN VBG DT NN IN NN JJ NN IN NNS IN DT NN .

NP Chunking

Look at the POS-tags

NN NN NNS VBP RB VBN TO VB JJ NN IN DT JJR JJ NN . IN VBN JJ JJ NN (FW) NNS IN CD CC JJR NNS) IN JJ NNS , in PDT DT NN IN JJ NNS VBP VBD IN JJ NNS CC JJ NNS VBP VBD IN JJ NNS , NNS MD VB NN IN VBG DT NN IN NN JJ NN IN NNS IN DT NN .

Approach 1

- Write Regular-expressions over POS-tags.

NP Chunking

Look at the POS-tags

NN NN NNS VBP RB VBN TO VB JJ NN IN DT JJR JJ NN . IN VBN JJ JJ NN (FW) NNS IN CD CC JJR NNS) IN JJ NNS , in PDT DT NN IN JJ NNS VBP VBD IN JJ NNS CC JJ NNS VBP VBD IN JJ NNS , NNS MD VB NN IN VBG DT NN IN NN JJ NN IN NNS IN DT NN .

Approach 1

- Write Regular-expressions over POS-tags.

Approach 2

- Memorize POS-tag patterns.
- Match the longest pattern.

NP Chunking

Approach 1

- Write Regular-expressions over POS-tags.

Approach 2

- Memorize POS-tag patterns.
- Match the longest pattern.

Downsides

- 1 is hard to maintain.
- 2 will cover the common cases, but may not generalize.

NP Chunking

Approach 1

- Write Regular-expressions over POS-tags.

Approach 2

- Memorize POS-tag patterns.
- Match the longest pattern.

Downsides

- 1 is hard to maintain.
- 2 will cover the common cases, but may not generalize.
- What if POS-tagger is wrong?

NP Chunking

Approach 1

- Write Regular-expressions over POS-tags.

Approach 2

- Memorize POS-tag patterns.
- Match the longest pattern.

Downsides

- 1 is hard to maintain.
- 2 will cover the common cases, but may not generalize.
- What if POS-tagger is wrong?
- What about other tasks?

Sequence Segmentation – Methods

need a learning approach

Sequence Segmentation – Learning

[Data visualization tools] are often used to communicate [complex information] in [a more intuitive way] . By representing [high dimensional data] (e.g. [objects] with [4 or more variables]) by [two-dimensional points] , in [such a way] that [similar objects] are represented by [nearby points] and [dissimilar objects] are represented by [distant points] , [scientists] can gain [intuition] by examining [the presence] of [structure and clustering] in [plots] of [the data] .

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`

Sequence Segmentation – Learning

[Data visualization tools] are often used to communicate [complex information] in [a more intuitive way] . By representing [high dimensional data] (e.g. [objects] with [4 or more variables]) by [two-dimensional points] , in [such a way] that [similar objects] are represented by [nearby points] and [dissimilar objects] are represented by [distant points] , [scientists] can gain [intuition] by examining [the presence] of [structure and clustering] in [plots] of [the data] .

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Sequence Segmentation – Learning

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Example

? Data visualization tools are often used to communicate complex information in a more intuitive way . By representing

$$\underset{y \in \{\texttt{]}, \texttt{[}, \texttt{][}, \texttt{None}\}}{\arg\max} \; w \cdot \phi(\text{-START-, -START-, ? Data/NN, visualization/JJ}, y)$$

Sequence Segmentation – Learning

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Example

Data **?** visualization tools are often used to communicate complex information in a more intuitive way . By representing

$$\underset{y \in \{\texttt{]},\texttt{[},\texttt{][},\texttt{None}\}}{\arg\max} \; w \cdot \phi(\text{-START, Data/NN, ?, visualization/JJ, tools/NNS}, y)$$

Sequence Segmentation – Learning

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Example

Data visualization **?** tools are often used to communicate complex information in a more intuitive way . By representing

$$\underset{y \in \{\texttt{]},\texttt{[},\texttt{][},\texttt{None}\}}{\arg\max}\; w \cdot \phi(\text{Data/NN, visualization/JJ, ? tools/NNS, are/VBP}, y)$$

Sequence Segmentation – Learning

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Example

Data visualization tools **?** are often used to communicate complex information in a more intuitive way . By representing

$$\underset{y \in \{\texttt{]},\texttt{[},\texttt{][},\texttt{None}\}}{\arg\max}\ w \cdot \phi(\text{visualization/JJ, tools/NNS, ?, are/VBP, often/RB}, y)$$

Sequence Segmentation – Learning

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Example

Data visualization tools are **?** often used to communicate complex information in a more intuitive way . By representing

$$\underset{y \in \{\texttt{]},\texttt{[},\texttt{][},\texttt{None}\}}{\arg\max}\ w \cdot \phi(\text{tools/NNS, are/VBP, ?, often/RB, used/VBD}, y)$$

Sequence Segmentation – Learning

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Example

Data visualization tools are often used to communicate complex information in a more intuitive way . By representing

$$\underset{y \in \{\texttt{]}, \texttt{[}, \texttt{][}, \texttt{None}\}}{\arg\max} \; w \cdot \phi(\underbrace{\text{tools/NNS, are/VBP, ?, often/RB, used/VBD}}_{\text{Sliding Window}}, y)$$

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `] [`

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

$n * 4$

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

$n * 4$ or, with span-types: $n * (2 + 2 * \text{num\_types})$

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

$n * 4$ or, with span-types: $n * (2 + 2 * \text{num\_types})$

Problems?

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

$n * 4$ or, with span-types: $n * (2 + 2 * \text{num\_types})$

Problems?

- Capturing the structures we need

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

$n * 4$ or, with span-types: $n * (2 + 2 * \text{num\_types})$

Problems?

- Capturing the structures we need
  - `[ A [ B ][ C ] D E ][ F ] G ][`

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

$n * 4$ or, with span-types: $n * (2 + 2 * \text{num\_types})$

Problems?

- Capturing the structures we need
  - `[ A [ B ][ C ] D E ][ F ] G ][`
- Modeling / Learning

Suggestion – Open / Close

- Classify each space between words in sequence as:
  - `[-type`
  - `]`
  - `None`
  - `][-type`

Complexity?

$n * 4$ or, with span-types: $n * (2 + 2 * \text{num\_types})$

Problems?

- Capturing the structures we need
  - `[ A [ B ][ C ] D E ][ F ] G ][`
- Modeling / Learning
  - inside and outside words have the same class

The Elegant/Correct Solution

- Assign a score to each possible span:

$\Rightarrow$ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

Data visualization tools are often used to communicate complex information in a more intuitive way . By representing

The Elegant/Correct Solution

- Assign a score to each possible span:

⇒ $score(i,j,type)$ – score of having span of type $type$ from word i to word j.

$\overbrace{\text{Data visualization tools}}^{score(1,3,NP)}$ are often used to communicate complex information in a more intuitive way . By representing

The Elegant/Correct Solution

- Assign a score to each possible span:

$\Rightarrow$ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

$\overbrace{\text{Data visualization tools}}^{score(1,3,VP)}$ are often used to communicate complex information in a more intuitive way . By representing

The Elegant/Correct Solution

- Assign a score to each possible span:

$\Rightarrow$ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

Data $\overbrace{\text{visualization tools}}^{score(2,3,NP)}$ are often used to communicate complex information in a more intuitive way . By representing

The Elegant/Correct Solution

- Assign a score to each possible span:

$\Rightarrow$ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

Data visualization $\overbrace{\text{tools}}^{score(3,3,NP)}$ are often used to communicate complex information in a more intuitive way . By representing

The Elegant/Correct Solution

- Assign a score to each possible span:

⇒ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

Data visualization $\overbrace{\text{tools are often used to}}^{score(3,7,NP)}$ communicate complex information in a more intuitive way . By representing

The Elegant/Correct Solution

- Assign a score to each possible span:

$\Rightarrow$ $score(i,j,type)$ – score of having span of type $type$ from word i to word j.

Data visualization $\overbrace{\text{tools are often used to}}^{score(3,7,NP)}$ communicate complex information in a more intuitive way . By representing

$$score(i,j,type) = w \cdot \phi(i,j,type,s_1,\ldots,s_n,t_1,\ldots,t_n)$$

- $s_1,\ldots,s_n$: the input sequence.
- $t_1,\ldots,t_n$: the POS-tags of the input sequence.

The Elegant/Correct Solution

- Assign a score to each possible span:

$\Rightarrow$ $score(i,j,type)$ – score of having span of type $type$ from word i to word j.

Data visualization $\overbrace{\text{tools are often used to}}^{score(3,7,NP)}$ communicate complex information in a more intuitive way . By representing

$$score(i,j,type) = w \cdot \phi(i,j,type,s_1,\ldots,s_n,t_1,\ldots,t_n)$$

- Learn w such that:
  - good spans have positive scores
  - bad spans have negative scores

The Elegant/Correct Solution

- Assign a score to each possible span:

$\Rightarrow$ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

Data visualization $\overbrace{\text{tools are often used to}}^{score(3,7,NP)}$ communicate complex information in a more intuitive way . By representing

$$score(i, j, type) = w \cdot \phi(i, j, type, s_1, \ldots, s_n, t_1, \ldots, t_n)$$

- Learn w such that:
  - good spans have positive scores
  - bad spans have negative scores
- Possible features?
  - discuss

The Elegant/Correct Solution

$\Rightarrow$ $score(i,j,type)$ – score of having span of type $type$ from word i to word j.

$$score(i,j,type) = w \cdot \phi(i,j,type,s_1,\ldots,s_n,t_1,\ldots,t_n)$$

- a segmentation y is a set of **consistent** triplets $(i,j,type)$

The Elegant/Correct Solution

⇒ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

$$score(i, j, type) = w \cdot \phi(i, j, type, s_1, \ldots, s_n, t_1, \ldots, t_n)$$

- a segmentation y is a set of **consistent** triplets $(i, j, type)$
  - ⇒ $1 \leq i \leq j \leq n$
  - ⇒ No overlaps

The Elegant/Correct Solution

⇒ $score(i,j,type)$ – score of having span of type $type$ from word i to word j.

$$score(i,j,type) = w \cdot \phi(i,j,type,s_1,\ldots,s_n,t_1,\ldots,t_n)$$

- a segmentation y is a set of **consistent** triplets $(i,j,type)$
  - ⇒ $1 \leq i \leq j \leq n$
  - ⇒ No overlaps
- The set of all segmentations: $\mathcal{Y}$

The Elegant/Correct Solution

$\Rightarrow$ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

$$score(i, j, type) = w \cdot \phi(i, j, type, s_1, \ldots, s_n, t_1, \ldots, t_n)$$

- a segmentation y is a set of **consistent** triplets $(i, j, type)$
  - $\Rightarrow$ $1 \leq i \leq j \leq n$
  - $\Rightarrow$ No overlaps
- The set of all segmentations: $\mathcal{Y}$
- We want:

$$\arg\max_{y \in \mathcal{Y}} \sum_{(i,j,type) \in y} score(i, j, type)$$

The Elegant/Correct Solution

⇒ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

$$score(i, j, type) = w \cdot \phi(i, j, type, s_1, \ldots, s_n, t_1, \ldots, t_n)$$

- a segmentation y is a set of **consistent** triplets $(i, j, type)$
  - ⇒ $1 \leq i \leq j \leq n$
  - ⇒ No overlaps
- The set of all segmentations: $\mathcal{Y}$
- We want:

$$\arg\max_{y \in \mathcal{Y}} \sum_{(i,j,type) \in y} score(i, j, type)$$

- Can solve with a dynamic program

The Elegant/Correct Solution

⇒ $score(i, j, type)$ – score of having span of type $type$ from word i to word j.

$$score(i, j, type) = w \cdot \phi(i, j, type, s_1, \ldots, s_n, t_1, \ldots, t_n)$$

- a segmentation y is a set of **consistent** triplets $(i, j, type)$
  - ⇒ $1 \leq i \leq j \leq n$
  - ⇒ No overlaps
- The set of all segmentations: $\mathcal{Y}$
- We want:

$$\arg\max_{y \in \mathcal{Y}} \sum_{(i,j,type) \in y} score(i, j, type)$$

- Can solve with a dynamic program
- This approach is called "semi-markov"

The Elegant/Correct Solution

- a segmentation y is a set of **consistent** triplets $(i, j, type)$
- The set of all segmentations: $\mathcal{Y}$
- We search for:

$$\arg\max_{y \in \mathcal{Y}} \sum_{(i,j,type) \in y} score(i, j, type)$$

$$score(i, j, type) = w \cdot \phi(i, j, type, s_1, \ldots, s_n, t_1, \ldots, t_n)$$

Training

Find w such that, for a correct segmentation y,
$\forall y' \in \mathcal{Y}, y' \neq y$:

$$\sum_{(i,j,type) \in y} score(i, j, type) > \sum_{(i',j',type') \in y'} score(i', j', type')$$

The Elegant/Correct Solution

- Assume we can train w.
- When we receive a new sequence $s_1, \ldots, s_n$, we need to solve:

$$\arg\max_{y \in \mathcal{Y}} \sum_{(i,j,type) \in y} score(i,j,type)$$

Complexity?

The Elegant/Correct Solution

- Assume we can train w.
- When we receive a new sequence $s_1, \ldots, s_n$, we need to solve:

$$\arg\max_{y \in \mathcal{Y}} \sum_{(i,j,type) \in y} score(i, j, type)$$

Complexity?

- We need to compute $score(i, j, type)$
  for all $1 \leq i \leq j \leq n$
  for all $type$.

$\Rightarrow$ Complexity is at least $O(kn^2)$, where k is the number of types.

The Elegant/Correct Solution

- Assume we can train w.
- When we receive a new sequence $s_1, \ldots, s_n$, we need to solve:

$$\arg\max_{y \in \mathcal{Y}} \sum_{(i,j,type) \in y} score(i,j,type)$$

Complexity?

- We need to compute $score(i,j,type)$
  for all $1 \leq i \leq j \leq n$
  for all $type$.

$\Rightarrow$ Complexity is at least $O(kn^2)$, where k is the number of types.

- Not cheap.

a different approach

Reminder

Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Complexity?

$(2k+2)n$

Reminder

Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `] [`

Complexity?

$(2k+2)n$ Cheap!

Reminder

Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Complexity?

$(2k+2)n$ Cheap!

Problems?

Reminder

Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Complexity?

$(2k+2)n$ Cheap!

Problems?

- Capturing the structures we need

Reminder

Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Complexity?

$(2k+2)n$ Cheap!

Problems?

- Capturing the structures we need
  - `[ A [ B ][ C ] D E ][ F ] G ][`

Reminder

Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `] [`

Complexity?

$(2k+2)n$ Cheap!

Problems?

- Capturing the structures we need
  - `[ A [ B ] [ C ] D E ] [ F ] G ] [`
- Modeling / Learning

Reminder

Open / Close

- Classify each space between words in sequence as:
  - `[`
  - `]`
  - `None`
  - `][`

Complexity?

$(2k+2)n$ Cheap!

Problems?

- Capturing the structures we need
  - `[ A [ B ][ C ] D E ][ F ] G ][`
- Modeling / Learning
  - inside and outside words have the same class

Sequence Segmentation – IO encoding

Suggestion – Inside / Out

- Classify each word in sequence as:
  - **I**nside a segment
  - **O**ut of segment

Data/I-NP visualization/I-NP tools/I-NP are/O often/O used/O to/O communicate/O complex/I-NP information/I-NP in/O a/I-NP more/I-NP intuitive/I-NP way/I-NP ./O By/O representing/O

Sequence Segmentation – IO encoding

Suggestion – Inside / Out

- Classify each word in sequence as:
  - **I**nside a segment
  - **O**ut of segment

Data/I-NP visualization/I-NP tools/I-NP are/O often/O used/O to/O communicate/O complex/I-NP information/I-NP in/O a/I-NP more/I-NP intuitive/I-NP way/I-NP ./O By/O representing/O

What did we gain?

Sequence Segmentation – IO encoding

Suggestion – Inside / Out

- Classify each word in sequence as:
  - **I**nside a segment
  - **O**ut of segment

Data/I-NP visualization/I-NP tools/I-NP are/O often/O used/O to/O communicate/O complex/I-NP information/I-NP in/O a/I-NP more/I-NP intuitive/I-NP way/I-NP ./O By/O representing/O

What did we gain? Is this sufficient?

Sequence Segmentation – IO encoding

Suggestion – Inside / Out

- Classify each word in sequence as:
  - **I**nside a segment
  - **O**ut of segment

Data/I-NP visualization/I-NP tools/I-NP are/O often/O used/O to/O communicate/O complex/I-NP information/I-NP in/O a/I-NP more/I-NP intuitive/I-NP way/I-NP ./O By/O representing/O

What did we gain? Is this sufficient?

$\overbrace{\text{he}}^{\text{NP}}$ saw $\overbrace{\text{a blue tiger}}^{\text{NP}}$ $\overbrace{\text{yesterday}}^{\text{NP}}$

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment

Does this solve the previous problems?

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment

Does this solve the previous problems?

B-NP	O	B-NP	I-NP	I-NP	B-NP
he	saw	a	blue	tiger	yesterday

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment

Does this solve the previous problems?

B-NP	O	B-NP	I-NP	I-NP	B-NP
he	saw	a	blue	tiger	yesterday

Further elaboration

- add **S**ingle class, for a single-word segment
- add **E**nd class, for last-word of segment
  - why?

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment

Does this solve the previous problems?

S-NP	O	B-NP	I-NP	E-NP	S-NP
he	saw	a	blue	tiger	yesterday

Further elaboration

- add **S**ingle class, for a single-word segment
- add **E**nd class, for last-word of segment
  - why?

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out (/ Single / End)

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out (/ Single / End)

- **Classify each word in sequence** as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out (/ Single / End)

- **Classify each word in sequence** as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

we reduced sequence segmentation to sequence tagging

Sequence Segmentation – BIO encoding

Suggestion – Begin / Inside / Out (/ Single / End)

- **Classify each word in sequence** as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

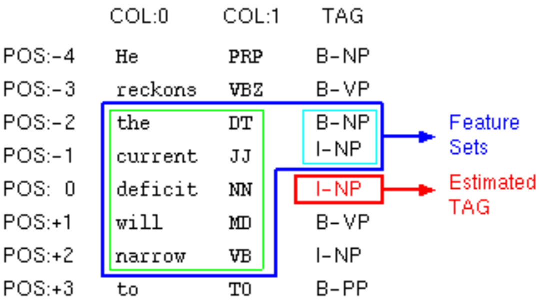
we reduced sequence segmentation to sequence tagging

we know how to train a tagger

Sequence Segmentation – BIO encoding

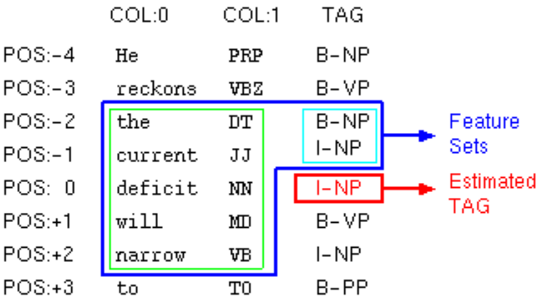
Suggestion – Begin / Inside / Out (/ Single / End)

- **Classify each word in sequence** as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)



Sequence Segmentation – BIO encoding

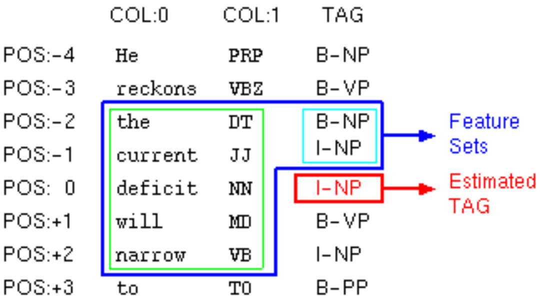
we reduced sequence segmentation to sequence tagging



- What did we gain over the “semi-markov” approach?

Sequence Segmentation – BIO encoding

we reduced sequence segmentation
to sequence tagging



- What did we gain over the “semi-markov” approach?
- What did we lose?

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`
  - `BOOOBOOOIIBBO`

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`
  - `BOOOBOOOIIBBO`
  - `BOOOIOOOIIBBO`

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`
  - `BOOOBOOOIIBBO`
  - `BOOOIOOOIIBBO`
  - `SOOOIOOOIEBSO`

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`
  - `BOOOBOOOIIBBO`
  - `BOOOIOOOIIBBO`
  - `SOOOIOOOIEBSO`
  - `SOOOEOOOIEBSO`

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`
  - `BOOOBOOOIIBBO`
  - `BOOOIOOOIIBBO`
  - `SOOOIOOOIEBSO`
  - `SOOOEOOOIEBSO`
- No good solution, just be aware of this

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`
  - `BOOOBOOOIIBBO`
  - `BOOOIOOOIIBBO`
  - `SOOOIOOOIEBSO`
  - `SOOOEOOOIEBSO`
- No good solution, just be aware of this
- When training, choose one.

Sequence Segmentation – BIO encoding

- Classify each word in sequence as:
  - **B**egin of segment
  - **I**nside a segment
  - **O**ut of segment
  - (**E**nd of segment)
  - (**S**ingle-word segment)

Notice: spurious ambiguity

- Several sequences encode the same structure:
  - `BOOOBOOOBIBBO`
  - `BOOOBOOOIIBBO`
  - `BOOOIOOOIIBBO`
  - `SOOOIOOOIEBSO`
  - `SOOOEOOOIEBSO`
- No good solution, just be aware of this
- When training, choose one.
- When predicting, know how to handle all.

A closer look at Named Entity Recognition

NER – Reminder

Named Entity Recognition (NER)

Paris Whitney Hilton (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Identify things

Person / **Loc**ation / **Org**anization / **Time** / **Other**.

NER – Reminder

Named Entity Recognition (NER)

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Named Entity Recognition

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981)}}^{\text{TIME}}$ is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York,}}^{\text{LOC}}$ $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Design questions

Assume we are using BIO tagging.

Named Entity Recognition

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Design questions

Assume we are using BIO tagging.

- BIO of BIOSE?

Named Entity Recognition

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Design questions

Assume we are using BIO tagging.

- BIO of BIOSE?
- Greedy or non-greedy tagger?

Named Entity Recognition

$\overbrace{\textbf{Paris Whitney}\text{ Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Design questions

Assume we are using BIO tagging.

- BIO of BIOSE?
- Greedy or non-greedy tagger?

Named Entity Recognition

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\textbf{Donald}\text{ Trump's modeling }\textbf{agency}}^{\text{ORG}}$.

Design questions

Assume we are using BIO tagging.

- BIO of BIOSE?
- Greedy or non-greedy tagger?

Named Entity Recognition

$\overbrace{\text{Paris Whitney Hilton}}^{\text{PER}}$ (born $\overbrace{\text{February 17, 1981}}^{\text{TIME}}$) is an American television personality and businesswoman . She is the great-granddaughter of $\overbrace{\text{Conrad Hilton}}^{\text{PER}}$, the founder of $\overbrace{\text{Hilton Hotels}}^{\text{ORG}}$. Born in $\overbrace{\text{New York City}}^{\text{LOC}}$ and raised in both $\overbrace{\text{California}}^{\text{LOC}}$ and $\overbrace{\text{New York}}^{\text{LOC}}$, $\overbrace{\text{Hilton}}^{\text{PER}}$ began a modeling career when she signed with $\overbrace{\text{Donald Trump's modeling agency}}^{\text{ORG}}$.

Design questions

Assume we are using BIO tagging.

- BIO of BIOSE?
- Greedy or non-greedy tagger?
- Features!!!

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.
- Is (current/prev/next) word capitalized?

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.
- Is (current/prev/next) word capitalized?
  - Should we rely on capitalization?

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.
- Is (current/prev/next) word capitalized?
  - Should we rely on capitalization?
  - Maybe: is the word usually capitalized, in a large corpus?

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.
- Is (current/prev/next) word capitalized?
  - Should we rely on capitalization?
  - Maybe: is the word usually capitalized, in a large corpus?
- Gazetteers (lists): is word in list of persons / locations / organizations?

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of Conrad Hilton, the founder of Hilton Hotels. Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.
- Is (current/prev/next) word capitalized?
  - Should we rely on capitalization?
  - Maybe: is the word usually capitalized, in a large corpus?
- Gazetteers (lists): is word in list of persons / locations / organizations?
  - Where do we find gazetteers?
  - How do we match against them?

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of **Conrad** Hilton, **the** founder **of** Hilton **Hotels**. Born in New York City and raised in both California and New **York**, Hilton **began** a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.
- Is (current/prev/next) word capitalized?
  - Should we rely on capitalization?
  - Maybe: is the word usually capitalized, in a large corpus?
- Gazetteers (lists): is word in list of persons / locations / organizations?
  - Where do we find gazetteers?
  - How do we match against them?
- Context-aggregation (prev/next words of other occurrences of current word)

Named Entity Recognition

Paris Whitney Hilton/**??** (born February 17, 1981) is an American television personality and businesswoman. She is the great-granddaughter of **Conrad** Hilton, the **founder** of Hilton **Hotels**. Born in New York City and raised in both California and New **York**, Hilton **began** a modeling career when she signed with Donald Trump's modeling agency .

Features

- Current word, previous word, next word.
- Is (current/prev/next) word capitalized?
  - Should we rely on capitalization?
  - Maybe: is the word usually capitalized, in a large corpus?
- Gazetteers (lists): is word in list of persons / locations / organizations?
  - Where do we find gazetteers?
  - How do we match against them?
- Context-aggregation (prev/next words of other occurrences of current word)
  - maybe better: prev/next non-function-word?

Named Entity Recognition

Design questions

- Features!!!
- Separate identification from classification?

Named Entity Recognition

Design questions

- Features!!!
- Separate identification from classification?

1) Paris Whitney Hilton (born February 17, 1981) is an American television personality and businesswoman . She is the great-granddaughter of Conrad Hilton , the founder of Hilton Hotels . Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

Named Entity Recognition

Design questions

- Features!!!
- Separate identification from classification?

1) Paris Whitney Hilton (born February 17, 1981) is an American television personality and businesswoman . She is the great-granddaughter of Conrad Hilton , the founder of Hilton Hotels . Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

2)
Paris Whitney Hilton $\rightarrow$ PER / LOC / ORG / TIME ?
February 17, 1981
Conrad Hilton
Hilton Hotels
New York City
California
New York
Hilton
Donald Trump's modeling agency

Named Entity Recognition

Design questions

- Features!!!
- Separate identification from classification?

1) Paris Whitney Hilton (born February 17, 1981) is an American television personality and businesswoman . She is the great-granddaughter of Conrad Hilton , the founder of Hilton Hotels . Born in New York City and raised in both California and New York, Hilton began a modeling career when she signed with Donald Trump's modeling agency .

2)

Paris Whitney Hilton → PER / LOC / ORG / TIME ?
February 17, 1981
Conrad Hilton
Hilton Hotels
New York City
California
New York
Hilton
Donald Trump's modeling agency

Why?

Summary

Sequence Segmentation

- Many tasks can be cast as sequence segmentation
- Evaluation: Precision / Recall / F_β
- Solutions
  - Elegant: semi-markov model
  - Efficient: BIO(SE) tagging
    - Spurious ambiguity
- Features!!
  - gazetteers
  - context aggregation
- Separate identification from classification

  . . . when it makes sense