General sequence tagging: NER and Chunking

Jeremy Barnes
HAP/LAP Master

17.01.2022

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Quick review

Review: with your groups

1. For a trigram model, how do we refactor $p(w_1, w_2, \ldots, w_n)$?
2. For a bigram HMM, how do we refactor $p(x, y)$?
3. What parameters does an HMM have?
4. Annotate the following sentences with UD POS tags
    - The dog jumped on my back.
    - The politicians backed the proposal.
    - I want my book back.

Review: with your groups: answers

1. For a trigram model, how do we refactor $p(w_1, w_2, \ldots, w_n)$?
   - $p(w_1, w_2, \ldots, w_3) = \prod_{i=1}^{n} p(w_i | w_{i-2}, w_{i-1})$
2. For a bigram HMM, how do we refactor $p(x, y)$?
   - $p(x, y) = p(y)p(x|y) = \prod_{i=1}^{n} p(y_i | y_{i-1}) \prod_{i=1}^{n} p(x_i | y_i)$
3. What parameters does an HMM have?
   - initialization probabilities: π
   - transition probabilities: A
   - emission probabilities: B
4. Annotate the following sentences with UD POS tags
   - The/DET dog/NOUN jumped/VERB on/ADP my/PRON back/NOUN.
   - The/DET politicians/NOUN backed/VERB the/DET proposal/NOUN.
   - I/PRON want/VERB my/PRON book/NOUN back/ADV.

Tagging sequences

Until now

- Language modeling: $p(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} p(w_i | w_{i-1})$
- POS tagging: $p(x, y) = p(y)p(x|y) = \prod_{i=1}^{n} p(y_i | y_{i-1}) \prod_{i=1}^{n} p(x_i | y_i)$
- Notice that we don't really have any constraints on what y is.
- So far, we have seen how to parameterize the model when y is part of speech tags.
- Today we'll see how to apply the same model to other tasks.

Information extraction

Information extraction

Motivation

- The internet is currently mainly a collection of unstructured data.

Information extraction

Motivation

- The internet is currently mainly a collection of unstructured data.
- Not easy to retrieve most information in a useful form.

Motivation

- The internet is currently mainly a collection of unstructured data.
- Not easy to retrieve most information in a useful form.
- Instead, we could process this information and keep it in another format:

Information extraction

Motivation

- The internet is currently mainly a collection of unstructured data.
- Not easy to retrieve most information in a useful form.
- Instead, we could process this information and keep it in another format:
  - Should be easily machine readable:
  - Relational database
  - XML markup

Information extraction

Given a text

Information extraction

Given a text

1. Find all the entities in the text.

Information extraction

Given a text

1. Find all the entities in the text.
2. Perform co-reference resolution.

Information extraction

Given a text

1. Find all the entities in the text.
2. Perform co-reference resolution.
3. Determine what relationship they have.

Information extraction

Given a text

1. Find all the entities in the text.
2. Perform co-reference resolution.
3. Determine what relationship they have.
4. Use this information to populate the database

Information extraction

Uses

Information extraction

Uses

- News paper articles

Information extraction

Uses

- News paper articles
- Web pages

Information extraction

Uses

- News paper articles
- Web pages
- Scientific articles

Information extraction

Uses

- News paper articles
- Web pages
- Scientific articles
- Medical notes

Knowledge Base Population

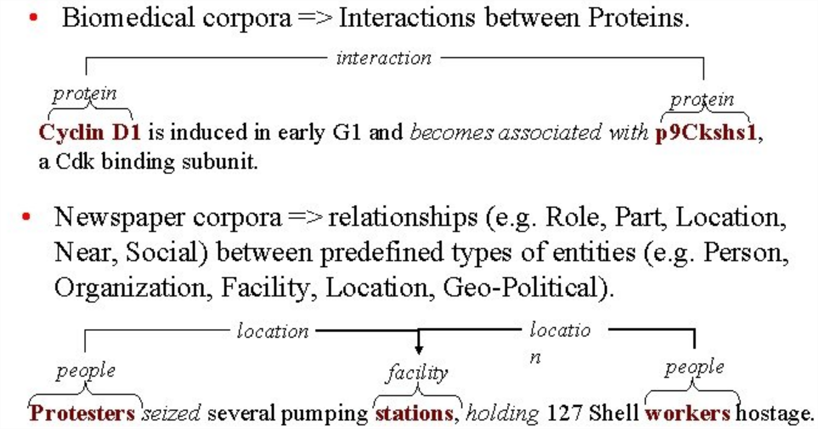
Penner is survived by his brother, John, a copy editor at the Times, and his former wife, Times sportswriter Lisa Dillman.

Knowledge Base Population

Penner is survived by his brother, John, a copy editor at the Times, and his former wife, Times sportswriter Lisa Dillman.

Subject	Relation/Slot	Object
Mike Penner	per:spouse	Lisa Dillman
Lisa Dillman	per:title	Sportswriter
Lisa Dillman	per:employee_of	Los Angeles Times
…	…	…

Information extraction



Named Entity Recognition

Named Entity Recognition

Definition

Named Entity Recognition

Definition

- First step in the pipeline.

Named Entity Recognition

Definition

- First step in the pipeline.
- The actual tags depend highly on the final use case.

Named Entity Recognition

Definition

- First step in the pipeline.
- The actual tags depend highly on the final use case.
- In research, we often use data from CONLL shared task:

Named Entity Recognition

Definition

- First step in the pipeline.
- The actual tags depend highly on the final use case.
- In research, we often use data from CONLL shared task:
  - PER: person
  - LOC: location
  - ORG: organization
  - MISC: miscellaneous

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Differences

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Differences

- There are two major differences between NER and POS tagging that we need to deal with before we can apply the HMM from last class.
- Can you see the two main differences?

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Differences

- There are two major differences between NER and POS tagging that we need to deal with before we can apply the HMM from last class.
- Can you see the two main differences?
  - Labels can span across several tokens.
  - A token doesn't have to have a label.

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Differences

- There are two major differences between NER and POS tagging that we need to deal with before we can apply the HMM from last class.
- Can you see the two main differences?
  - Labels can span across several tokens.
  - A token doesn't have to have a label.

Any ideas on how we could deal with these?

(With partners for 5 minutes)

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Options

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Options

- Since we already have a model we really like, we can change the label structure

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Options

- Since we already have a model we really like, we can change the label structure
- "If all you have is a hammer, everything looks like a nail" - Abraham Maslow

Named Entity Recognition

Options

- Since we already have a model we really like, we can change the label structure
- "If all you have is a hammer, everything looks like a nail" - Abraham Maslow
- Alternatively, "If you have a good hammer, why make a new hammer for every problem?" - Unknown author

Named Entity Recognition

[PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .

Options

- 1. Let's make a small change and give all tokens a tag.

Named Entity Recognition

Wolff/PER , currently/O a/O journalist/O in/O Argentina/LOC ,/O played/O with/O Del/PER Bosque/PER in/O the/O final/O years/O of/O the/O seventies/O in/O Real/ORG Madrid/ORG ./O

Options

- 1. Let's make a small change and give all tokens a tag.
- Give the tokens not in a label the O tag: outside of label.
- But now it is not possible to ensure that tags that span over 2 or more tokens are retrievable.

Named Entity Recognition

Wolff/B-PER , currently/O a/O journalist/O in/O Argentina/B-LOC ,/O played/O with/O Del/B-PER Bosque/I-PER in/O the/O final/O years/O of/O the/O seventies/O in/O Real/B-ORG Madrid/I-ORG ./O

Options

- 1. Let's make a small change and give all tokens a tag.
- Give the tokens not in a label the O tag: outside of label.
- But now it is not possible to ensure that tags that span over 2 or more tokens are retrievable.
- Add beginning and inside to labels, i.e., B-ORG, I-ORG
- This tagging scheme is known as IOB or BIO.
- There are a few variants that propose improvements.

Chunking

Chunking

Robust, efficient approach to syntax

Chunking

Robust, efficient approach to syntax

- Also useful in information extraction
- Instead of full syntactic analysis
- Chunks:

Chunking

Robust, efficient approach to syntax

- Also useful in information extraction
- Instead of full syntactic analysis
- Chunks:
  - Non-recursive text spans
  - Includes a head and its modifiers

Wolff/PROPN ,/PUNCT currently/ADV a/DET journalist/NOUN in/ADP Argentina/NOUN ,/PUNCT played/VERB with/ADP Del/PROPN Bosque/PROPN in/ADP the/DET final/ADJ years/NOUN of/ADP the/DET seventies/NOUN in/ADP Real/PROPN Madrid/PROPN ./PUNCT

Chunking

Robust, efficient approach to syntax

- Also useful in information extraction
- Instead of full syntactic analysis
- Chunks:
  - Non-recursive text spans
  - Includes a head and its modifiers

[NP Wolff/PROPN] ,/PUNCT currently/ADV [NP a/DET journalist/NOUN in/ADP Argentina/NOUN] ,/PUNCT [VP played/VERB] with/ADP [NP Del/PROPN Bosque/PROPN] in/ADP [NP the/DET final/ADJ years/NOUN] of/ADP [NP the/DET seventies/NOUN] in/ADP [NP Real/PROPN Madrid/PROPN] ./PUNCT

Chunking

Quick exercise: Use IOB tagging for this chunking example

[NP Wolff] , currently [NP a journalist] in [NP Argentina] ,
[VP played] with [NP Del Bosque] in [NP the final years]
of [NP the seventies] in [NP Real Madrid] .

Chunking

Wolff	B-NP
,	O
currently	O
a	B-NP
journalist	I-NP
in	O
Argentina	B-NP
,	O
played	B-VP
with	O
Del	B-NP
Bosque	I-NP
in	O
the	B-NP
final	I-NP
years	I-NP
of	O
the	B-NP
seventies	I-NP
in	O
Real	B-NP
Madrid	I-NP
.	PUNCT

Evaluation of sequence tagging

Evaluation of sequence tagging

Let's go back to our NER example and think of how to evaluate this sentence.

Evaluation of sequence tagging

Wolff/B-PER , currently/O a/O journalist/O in/O Argentina/B-LOC ,/O played/O with/O Del/B-PER Bosque/I-PER in/O the/O final/O years/O of/O the/O seventies/O in/O Real/B-ORG Madrid/I-ORG ./O

Let's go back to our NER example and think of how to evaluate this sentence.

Evaluation of sequence tagging

Wolff/B-PER , currently/O a/O journalist/O in/O Argentina/B-LOC ,/O played/O with/O Del/B-PER Bosque/I-PER in/O the/O final/O years/O of/O the/O seventies/O in/O Real/B-ORG Madrid/I-ORG ./O

Let's go back to our NER example and think of how to evaluate this sentence.

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Evaluation of sequence tagging

Wolff/B-PER , currently/O a/O journalist/O in/O Argentina/B-LOC ,/O played/O with/O Del/B-PER Bosque/I-PER in/O the/O final/O years/O of/O the/O seventies/O in/O Real/B-ORG Madrid/I-ORG ./O

Let's go back to our NER example and think of how to evaluate this sentence.

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the accuracy?

Evaluation of sequence tagging

Wolff/B-PER , currently/O a/O journalist/O in/O Argentina/B-LOC ,/O played/O with/O Del/B-PER Bosque/I-PER in/O the/O final/O years/O of/O the/O seventies/O in/O Real/B-ORG Madrid/I-ORG ./O

Let's go back to our NER example and think of how to evaluate this sentence.

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the accuracy?

Acc = 0.90

Evaluation of sequence tagging

Ok, no problem, right?

Evaluation of sequence tagging

Ok, no problem, right?

A slightly different example

Evaluation of sequence tagging

Ok, no problem, right?

A slightly different example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Evaluation of sequence tagging

Ok, no problem, right?

A slightly different example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the accuracy?

Evaluation of sequence tagging

Ok, no problem, right?

A slightly different example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the accuracy?

Acc $= 0.80$

Evaluation of sequence tagging

Ok, no problem, right?

A slightly different example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the accuracy?

Acc = 0.80

So, if you just always predict 'O', you could easily achieve 80-90% accuracy.

Evaluation of sequence tagging

Ok, no problem, right?

A slightly different example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the accuracy?

Acc = 0.80

So, if you just always predict 'O', you could easily achieve 80-90% accuracy. Our artificially adding labels to all tokens means we can't really use accuracy anymore.

Evaluation of sequence tagging

What other metrics could we use?

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

- Precision: $\frac{\text{correct predictions}}{\text{all output predictions}}$

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

- Precision: $\frac{\text{correct predictions}}{\text{all output predictions}}$
- Recall: $\frac{\text{correct predictions}}{\text{all possible predictions}}$

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

- Precision: $\frac{\text{correct predictions}}{\text{all output predictions}}$
- Recall: $\frac{\text{correct predictions}}{\text{all possible predictions}}$
- F_1: $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$

Example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

- Precision: $\frac{\text{correct predictions}}{\text{all output predictions}}$
- Recall: $\frac{\text{correct predictions}}{\text{all possible predictions}}$
- F_1: $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$

Example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the precision, recall, F_1?

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

- Precision: $\frac{\text{correct predictions}}{\text{all output predictions}}$
- Recall: $\frac{\text{correct predictions}}{\text{all possible predictions}}$
- F_1: $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$

Example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the precision, recall, F_1?

Precision = 1.0 (Only one prediction, which was correct)

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

- Precision: $\frac{\text{correct predictions}}{\text{all output predictions}}$
- Recall: $\frac{\text{correct predictions}}{\text{all possible predictions}}$
- F_1: $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$

Example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the precision, recall, F_1?

Precision = 1.0 (Only one prediction, which was correct)

Recall = 0.5 (one of two)

Evaluation of sequence tagging

What other metrics could we use?

Other metrics

- Precision: $\frac{\text{correct predictions}}{\text{all output predictions}}$
- Recall: $\frac{\text{correct predictions}}{\text{all possible predictions}}$
- F_1: $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$

Example

gold =['B-PER', 'O', 'O', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O']

pred =['B-PER', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

What is the precision, recall, F_1?

Precision = 1.0 (Only one prediction, which was correct)

Recall = 0.5 (one of two)

$F_1 = \frac{2*1*0.5}{1+0.5} = 0.66$

Precision, Recall, and F_1 variants

Two main variants:

Precision, Recall, and F_1 variants

Two main variants:

- Micro-averaged and Macro-averaged

Precision, Recall, and F_1 variants

Two main variants:

- Micro-averaged and Macro-averaged
- Micro:
  - We count for all labels mixed
  - Used when you care about frequency of labels.

Precision, Recall, and F_1 variants

Two main variants:

- Micro-averaged and Macro-averaged
- Micro:
  - We count for all labels mixed
  - Used when you care about frequency of labels.
  - Prec: 1.0
  - Rec: 0.5
  - F_1: 0.66

Precision, Recall, and F_1 variants

Two main variants:

- Micro-averaged and Macro-averaged
- Micro:
  - We count for all labels mixed
  - Used when you care about frequency of labels.
  - Prec: 1.0
  - Rec: 0.5
  - F_1: 0.66
- Macro:

Precision, Recall, and F_1 variants

Two main variants:

- Micro-averaged and Macro-averaged
- Micro:
  - We count for all labels mixed
  - Used when you care about frequency of labels.
  - Prec: 1.0
  - Rec: 0.5
  - F_1: 0.66
- Macro:
  - We compute the metric for each label, and then average them
  - Used when you care equally about infrequent labels.

Precision, Recall, and F_1 variants

Two main variants:

- Micro-averaged and Macro-averaged
- Micro:
  - We count for all labels mixed
  - Used when you care about frequency of labels.
  - Prec: 1.0
  - Rec: 0.5
  - F_1: 0.66
- Macro:
  - We compute the metric for each label, and then average them
  - Used when you care equally about infrequent labels.
  - Prec: 0.5
  - Rec: 0.5
  - F_1: 0.5