Natural Language Processing 1

Lecture 8: Discourse processing

Katia Shutova

ILLC
University of Amsterdam

25 November 2019

Outline.

Discourse structure

Learning document representations

Referring expressions and coreference

Algorithms for coreference resolution

Document structure and discourse structure

- Most types of document are highly structured, implicitly or explicitly:
  - Scientific papers: conventional structure (differences between disciplines).
  - News stories: first sentence is a summary.
  - Blogs, etc etc
- Topics within documents.
- Relationships between sentences.

Rhetorical relations

Max fell. John pushed him.

can be interpreted as:

1. Max fell because John pushed him.
   EXPLANATION

or

2 Max fell and then John pushed him.
NARRATION

Implicit relationship: discourse relation or rhetorical relation
because, *and then* are examples of cue phrases

Rhetorical relations

Analysis of text with rhetorical relations generally gives a binary branching structure:

- nucleus (the main phrase) and satellite (the subsidiary phrase: e.g., EXPLANATION, JUSTIFICATION

  Max fell because John pushed him.

- equal weight: e.g., NARRATION

  Max fell and Kim kept running.

Rhetorical relations

Analysis of text with rhetorical relations generally gives a binary branching structure:

- nucleus (the main phrase) and satellite (the subsidiary phrase: e.g., EXPLANATION, JUSTIFICATION

  Max fell <u>because John pushed him</u>.

- equal weight: e.g., NARRATION

  Max fell and Kim kept running.

Coherence

Discourses have to have connectivity to be coherent:

Kim got into her car. Sandy likes apples.

Can be OK in context:

Kim got into her car. Sandy likes apples, so Kim thought she'd go to the farm shop and see if she could get some.

Coherence

Discourses have to have connectivity to be coherent:

Kim got into her car. Sandy likes apples.

Can be OK in context:

Kim got into her car. Sandy likes apples, so Kim thought she'd go to the farm shop and see if she could get some.

Coherence in interpretation

Discourse coherence assumptions can affect interpretation:

John likes Bill. He gave him an expensive Christmas present.

If EXPLANATION - ‘he’ is probably Bill.
If JUSTIFICATION (supplying evidence for another sentence), ‘he’ is John.

Factors influencing discourse interpretation

1. Cue phrases (e.g. *because, and*)
2. Punctuation (also prosody) and text structure.

   Max fell (John pushed him) and Kim laughed.
   Max fell, John pushed him and Kim laughed.
3. Real world content:

   Max fell. John pushed him as he lay on the ground.
4. Tense and aspect.

   Max fell. John had pushed him.
   Max was falling. John pushed him.

Discourse parsing

Discourse parsing: identifying discourse structure and relations

Hard problem, much research has focused on labelling relations between pairs of sentences / clauses

1. Classification with hand-engineered features
   - e.g. punctuation, cue phrases, syntactic and lexical
2. Neural models
   - take two sentences as input
   - train a sentence encoder
   - objective: predict the relation

Or learn document representations in a given task

Outline.

Discourse structure

Learning document representations

Referring expressions and coreference

Algorithms for coreference resolution

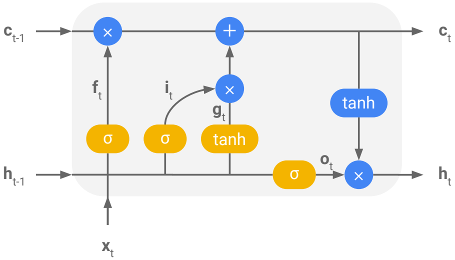
Document representations

Document classification tasks:

- text categorization (e.g. by topic)
- sentiment analysis
- authorship attribution
- spam and phishing email filtering
- misinformation detection
- and many more

Learning document representations

- Last time we have seen LSTMs for learning sentence representations

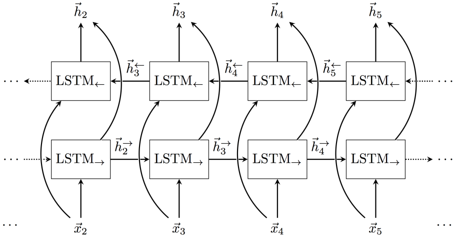


- Use these sentence representations to compute a document representation

Bidirectional LSTM

Bidirectional LSTM: BiLSTM

- Traverse the sentence in both directions



$$\overrightarrow{h_t} = \text{LSTM}^{\text{forward}}(\overrightarrow{h}_{t-1}, x_t)$$

$$\overleftarrow{h_t} = \text{LSTM}^{\text{backward}}(\overleftarrow{h}_{t+1}, x_t)$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$$

What is the sentence representation?

Options:

1. use h_L — the final hidden state of the LSTM
2. use an average of LSTM hidden states at all time steps (mean-pooling)
3. use max-pooling — take the maximum value in each vector component of all hidden states
4. use an attention mechanism, i.e. a weighted sum of the hidden states at all time steps

Attention mechanism

Sentence representation as a weighted sum of all hidden states

- the model learns a **weight vector** w_α, and computes its **dot product with the hidden state** h_t transformed by a FFNN:

$$\alpha_t = w_\alpha \cdot \mathrm{FFNN}_\alpha(h_t)$$

- normalise the weights into a distribution via softmax

$$a_t = \frac{e^{\alpha_t}}{\sum_{k=1}^{L} e^{\alpha_k}}$$

- compute the sentence representation h_{ATT} as a weighted sum

$$h_{ATT} = \sum_{t=1}^{L} a_t \cdot h_t$$

Building a document representation

Options:

1. Feed the whole document to an LSTM word by word
   - possibly use word-level attention to learn what are the useful words
2. Build a hierarchical model
   - first compute sentence representations
   - combine sentence representations into a document representation
   - using **another LSTM** and / or **attention over sentences**
   - train with a document level objective

Building a document representation

Options:

1. Feed the whole document to an LSTM word by word
   - possibly use word-level attention to learn what are the useful words
2. Build a hierarchical model
   - first compute sentence representations
   - combine sentence representations into a document representation
   - using **another LSTM** and / or **attention over sentences**
   - train with a document level objective

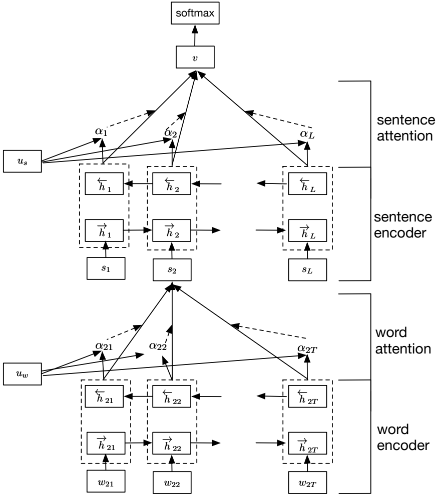
Hierarchical attention networks

Yang et al. 2016. *Hierarchical Attention Networks for Document Classification*. NAACL.

- Take pretrained word embeddings as input
- LSTM sentence encoder with word-level attention (to construct sentence representations)
- LSTM document encoder with sentence-level attention (to construct document representations)
- trained with document-level objective

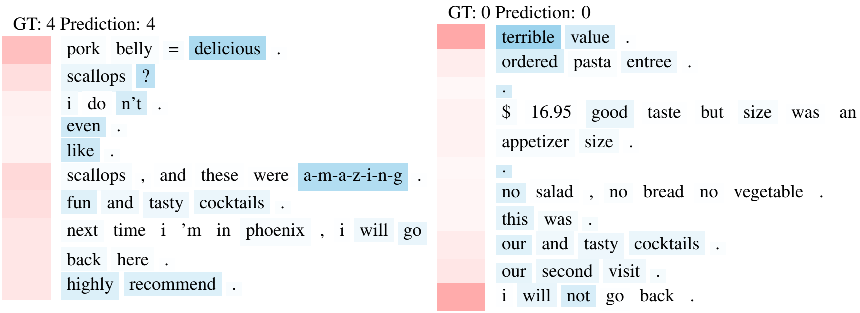
Experiments with sentiment analysis and text categorization

Hierarchical attention network



HAN output with attention visualised

Sentiment analysis of Yelp reviews



GT: 4 Prediction: 4

pork belly = delicious .
scallops ?
i do n't .
even .
like .
scallops , and these were a-m-a-z-i-n-g .
fun and tasty cocktails .
next time i 'm in phoenix , i will go back here .
highly recommend .

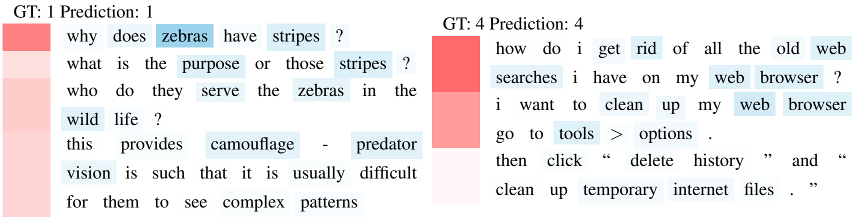
GT: 0 Prediction: 0

terrible value .
ordered pasta entree .
.
$ 16.95 good taste but size was an appetizer size .
.
no salad , no bread no vegetable .
this was .
our and tasty cocktails .
our second visit .
i will not go back .

Documents from Yelp 2013. Label 4 means star 5, label 0 means star 1.

HAN output with attention visualised

Topic classification



GT: 1 Prediction: 1

why does zebras have stripes ?
what is the purpose or those stripes ?
who do they serve the zebras in the wild life ?
this provides camouflage - predator vision is such that it is usually difficult for them to see complex patterns

GT: 4 Prediction: 4

how do i get rid of all the old web searches i have on my web browser ?
i want to clean up my web browser
go to tools > options .
then click " delete history " and " clean up temporary internet files . "

Documents from Yahoo Answers. Label 1 denotes Science and Mathematics and label 4 denotes Computers and Internet.

Outline.

Discourse structure

Learning document representations

Referring expressions and coreference

Algorithms for coreference resolution

Co-reference and referring expressions

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

referent a real world entity that some piece of text (or speech) refers to. the actual Prof. Ferguson

referring expressions bits of language used to perform reference by a speaker. 'Niall Ferguson', 'he', 'him'

antecedent the text initially evoking a referent. 'Niall Ferguson'

anaphora the phenomenon of referring to an antecedent.

cataphora pronouns appear before the referent (rare)

What about *a snappy dresser*?

Pronoun resolution

- Identifying the referents of pronouns
- Anaphora resolution: generally only consider cases which refer to antecedent noun phrases.

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

Pronoun resolution

- Identifying the referents of pronouns
- Anaphora resolution: generally only consider cases which refer to antecedent noun phrases.

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

Outline.

Discourse structure

Learning document representations

Referring expressions and coreference

Algorithms for coreference resolution

Coreference resolution as supervised classification

- instances: potential pronoun/antecedent pairings
- class is TRUE/FALSE
- training data labelled with correct pairings
- candidate antecedents are all NPs in current sentence and preceeding 5 sentences (excluding pleonastic pronouns)

Niall Ferguson is prolific, well-paid and a snappy dresser. Stephen Moss hated him — at least until he spent an hour being charmed in the historian's Oxford study.

Hard constraints: Pronoun agreement

- A little girl is at the door — see what she wants, please?
- My dog has hurt his foot — he is in a lot of pain.
- * My dog has hurt his foot — it is in a lot of pain.

Complications:

- I don’t know who the new lecturer will be, but I’m sure they’ll make changes to the course.
- The team played really well, but now they are all very tired.
- Kim and Sandy are asleep: they are very tired.

Hard constraints: Reflexives

- John$_i$ cut himself$_i$ shaving. (himself = John, subscript notation used to indicate this)
- # John$_i$ cut him$_j$ shaving. (i $\neq$ j — a very odd sentence)

Reflexive pronouns must be coreferential with a preceeding argument of the same verb, non-reflexive pronouns cannot be.

Hard constraints: Pleonastic pronouns

Pleonastic pronouns are semantically empty, and don't refer:

- It is snowing
- It is not easy to think of good examples.
- It is obvious that Kim snores.
- It bothers Sandy that Kim snores.

Soft preferences: Salience

- Recency: More recent antecedents are preferred. They are more accessible.

  Kim has a big car. Sandy has a smaller one. Lee likes to drive it.

- Grammatical role: Subjects $>$ objects $>$ everything else:

  Fred went to the shopping centre with Bill. He bought a CD.

- Repeated mention: Entities that have been mentioned more frequently are preferred.

Soft preferences: Salience

- Parallelism Entities which share the same role as the pronoun in the same sort of sentence are preferred:

  Bill went with Fred to the lecture. Kim went with him to the bar. Him=Fred

- Coherence effects: The pronoun resolution may depend on the rhetorical / discourse relation that is inferred.

  Bill likes Fred. He has a great sense of humour.

Features

Cataphoric Binary: t if pronoun before antecedent.

Number agreement Binary: t if pronoun compatible with antecedent.

Gender agreement Binary: t if gender agreement.

Same verb Binary: t if the pronoun and the candidate antecedent are arguments of the same verb.

Sentence distance Discrete: { 0, 1, 2 . . . }

Grammatical role Discrete: { subject, object, other } The role of the potential antecedent.

Parallel Binary: t if the potential antecedent and the pronoun share the same grammatical role.

Linguistic form Discrete: { proper, definite, indefinite, pronoun }

Feature vectors

Niall Ferguson is prolific, well-paid and a snappy dresser.
Stephen Moss hated him — at least until he spent an hour
being charmed in the historian's Oxford study.

pron	ante	cat	num	gen	same	dist	role	par	form
him	*Niall F.*	f	t	t	f	1	subj	f	prop
him	*Ste. M.*	f	t	t	t	0	subj	f	prop
him	*he*	t	t	t	f	0	subj	f	pron
he	*Niall F.*	f	t	t	f	1	subj	t	prop
he	*Ste. M.*	f	t	t	f	0	subj	t	prop
he	*him*	f	t	t	f	0	obj	f	pron

Apply any classifier, e.g. SVM, random forests etc.

Problems with simple classification model

- Cannot implement ‘repeated mention’ effect.
- Cannot use information from previous links.

Not really pairwise: need a discourse model with real world entities corresponding to clusters of referring expressions.

Neural end-to-end coreference resolution

Lee et al. 2017. *End-to-end Neural Coreference Resolution*. EMNLP.

- Mention-ranking paradigm, i.e. output a probability distribution over candidate mentions
- considers all text spans of certain length (e.g. bigrams, trigrams) as possible mentions
- coreference of all mentions considered (not only pronouns)
- end-to-end trainable neural architecture, based on an LSTM sentence encoder

Task definition

Assign each span i an antecedent y_i

- out of all possible spans in $Y_i = \{1, ..., i-1, \epsilon\}$
- empty token ϵ is included to indicate the span i is non-referential or discourse-new

To do this, for each pair of spans i and j

- the model **assigns a score** $s(i, j)$ for their coreference link
- and computes a distribution $P(y_i)$ over the antecedents of i

$$P(y_i) = \frac{e^{s(i,y_i)}}{\sum_{y' \in Y(i)} e^{s(i,y')}}$$

Computing the score s

The score $s(i, j)$ includes three factors:

- $m(i)$: whether span i is a mention
- $m(j)$: whether span j is a mention
- $c(i, j)$: whether j is the antecedent of i

$$s(i, j) = m(i) + m(j) + c(i, j)$$

$s(i, \epsilon)$ is set to 0, i.e. the model predicts the antecedent with the highest positive score or abstains

Computing the scoring functions m and c

- Compute $m(i)$, $m(j)$ and $c(i,j)$ based on the vectors g_i and g_j, which represent the spans i and j
- span representations are constructed from hidden states of the LSTM encoder:

$$g_i = [h_{START(i)}, h_{END(i)}, h_{ATT(i)}, \phi(i)],$$

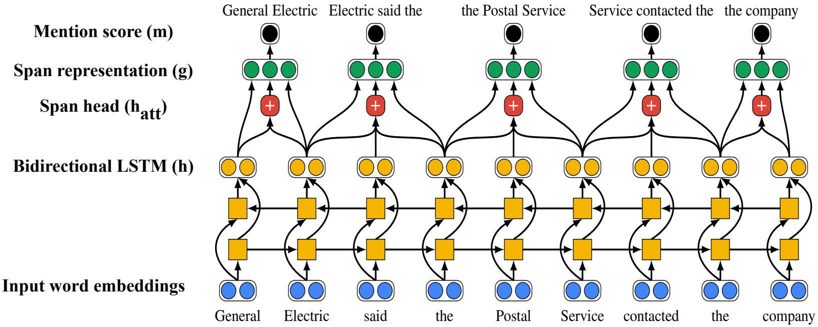
where $\phi(i)$ is a single feature: the length of the span

$$m(i) = w_m \cdot \text{FFNN}_m(g_i)$$

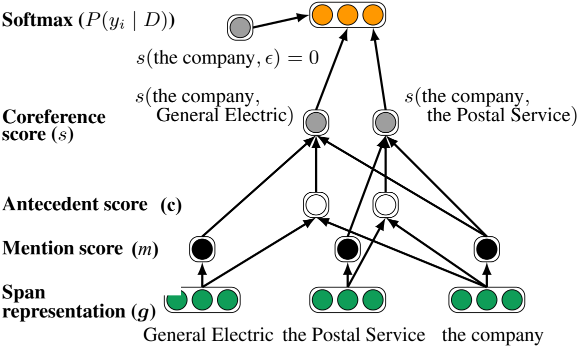
$$c(i,j) = w_c \cdot \text{FFNN}_c([g_i, g_j, g_i \odot g_j, \phi(i,j)])$$

$\phi(i,j)$ – distance between the spans in text

Learning span representations

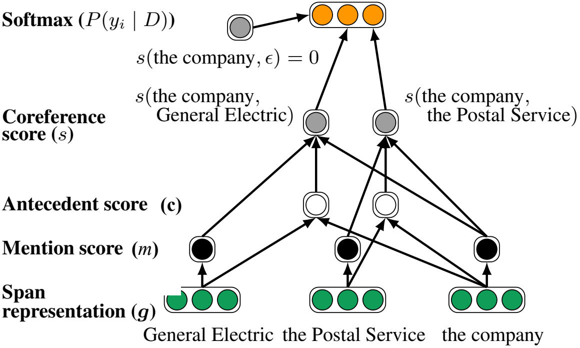


Computing the score


Softmax ($P(y_i \mid D)$)
$s(\text{the company}, \epsilon) = 0$
s(the company, General Electric)
s(the company, the Postal Service)
Coreference score (s)
Antecedent score (c)
Mention score (m)
Span representation (g)
General Electric
the Postal Service
the company

Train to maximize probabilities of valid mention pairings

Computing the score


Softmax ($P(y_i \mid D)$)
$s(\text{the company}, \epsilon) = 0$
s(the company, General Electric)
s(the company, the Postal Service)
Coreference score (s)
Antecedent score (c)
Mention score (m)
Span representation (g)
General Electric
the Postal Service
the company

Train to maximize probabilities of valid mention pairings

Model output with attention visualised

1	**(A fire in a Bangladeshi garment factory)** has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee **(the blaze)** in the four-story building.
	A fire in **(a Bangladeshi garment factory)** has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in **(the four-story building)**.
2	We are looking for **(a region of central Italy bordering the Adriatic Sea)**. **(The area)** is mostly mountainous and includes Mt. Corno, the highest peak of the Apennines. **(It)** also includes a lot of sheep, good clean-living, healthy sheep, and an Italian entrepreneur has an idea about how to make a little money of them.

Examples of errors

3	**(The flight attendants)** have until 6:00 today to ratify labor concessions. **(The pilots')** union and ground crew did so yesterday.
4	**(Prince Charles and his new wife Camilla)** have jumped across the pond and are touring the United States making **(their)** first stop today in New York. It's Charles' first opportunity to showcase his new wife, but few Americans seem to care. Here's Jeanie Mowth. What a difference two decades make. **(Charles and Diana)** visited a JC Penney's on the prince's last official US tour. Twenty years later here's the prince with his new wife.

Evaluation of coreference resolution

- link accuracy, i.e. percentage of correct links.
- link precision and recall

But:

- Identification of mentions should be part of the evaluation.
- Binary linkages don't allow for chains:

  Sally met Andrew in town and took him to the new restaurant. He was impressed.

Multiple evaluation metrics exist because of such problems.

Gender bias in coreference resolution

- WinoBias dataset (Zhao et al. 2018)
- test the gender bias of coreference resolution algorithms
- link gendered pronouns with [pro-/ anti-] stereotypical antecendents

  The secretary called the physician and told him about a new patient
  The secretary called the physician and told her about a new patient

- measure coreference resolution accuracy in pro- / anti-stereotypical cases.
- anti-stereotypical: F1-score 21% lower on average.

Gender bias in word embeddings

Zhao et al. 2017. *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*. EMNLP.

- Word embeddings encode gender stereotypes
- e.g. the following analogy from skip-gram

  man is to programmer as woman is to ?

- much research on debiasing them

Gender bias in word embeddings

Zhao et al. 2017. *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*. EMNLP.

- Word embeddings encode gender stereotypes
- e.g. the following analogy from skip-gram

  man is to programmer as woman is to homemaker

- much research on debiasing them

Appendix: LSTM formulae

hidden state

cell state

previous hidden state and cell state

$$\mathbf{h}_t, \mathbf{c}_t = \mathit{lstm}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1})$$

input gate $$\mathbf{i}_t = \sigma(W_i\,\mathbf{x}_t + R_i\,\mathbf{h}_{t-1} + \mathbf{b}_i)$$

forget gate $$\mathbf{f}_t = \sigma(W_f\,\mathbf{x}_t + R_f\,\mathbf{h}_{t-1} + \mathbf{b}_f)$$

candidate $$\mathbf{g}_t = \tanh(W_g\,\mathbf{x}_t + R_g\,\mathbf{h}_{t-1} + \mathbf{b}_g)$$

output gate $$\mathbf{o}_t = \sigma(W_o\,\mathbf{x}_t + R_o\,\mathbf{h}_{t-1} + \mathbf{b}_o)$$

cell state $$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

hidden state $$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

Acknowledgement

Some slides were adapted from Ann Copestake