Linguistic features

MMAI 5400 – lecture 6
Fall 2024

Contents

NLP tasks

There are many different NLP tasks.

NLP-progress: Tracking Progress in Natural Language Processing lists the tasks, descriptions & state-of-the-art (SOTA) results.

A few sentence-level tasks

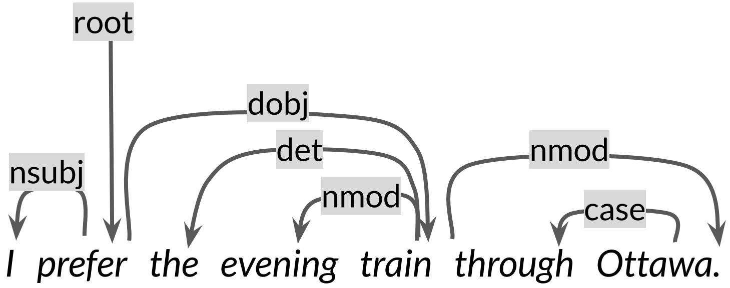
POS tagging

- Tag words with their part-of-speech (POS)

Coreference resolution

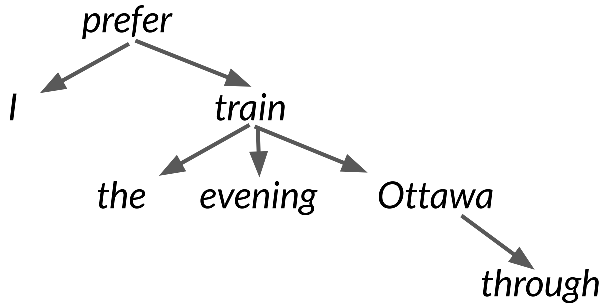
- Clustering mentions in text that refer to the same underlying real-world entities.
- Example: ***I*** *voted for* ***Trudeau*** *because* ***he*** *was most aligned with* ***my*** *values", **she** said.*
- http://nlpprogress.com/english/coreference_resolution.html

A few sentence-level tasks



Dependency parsing

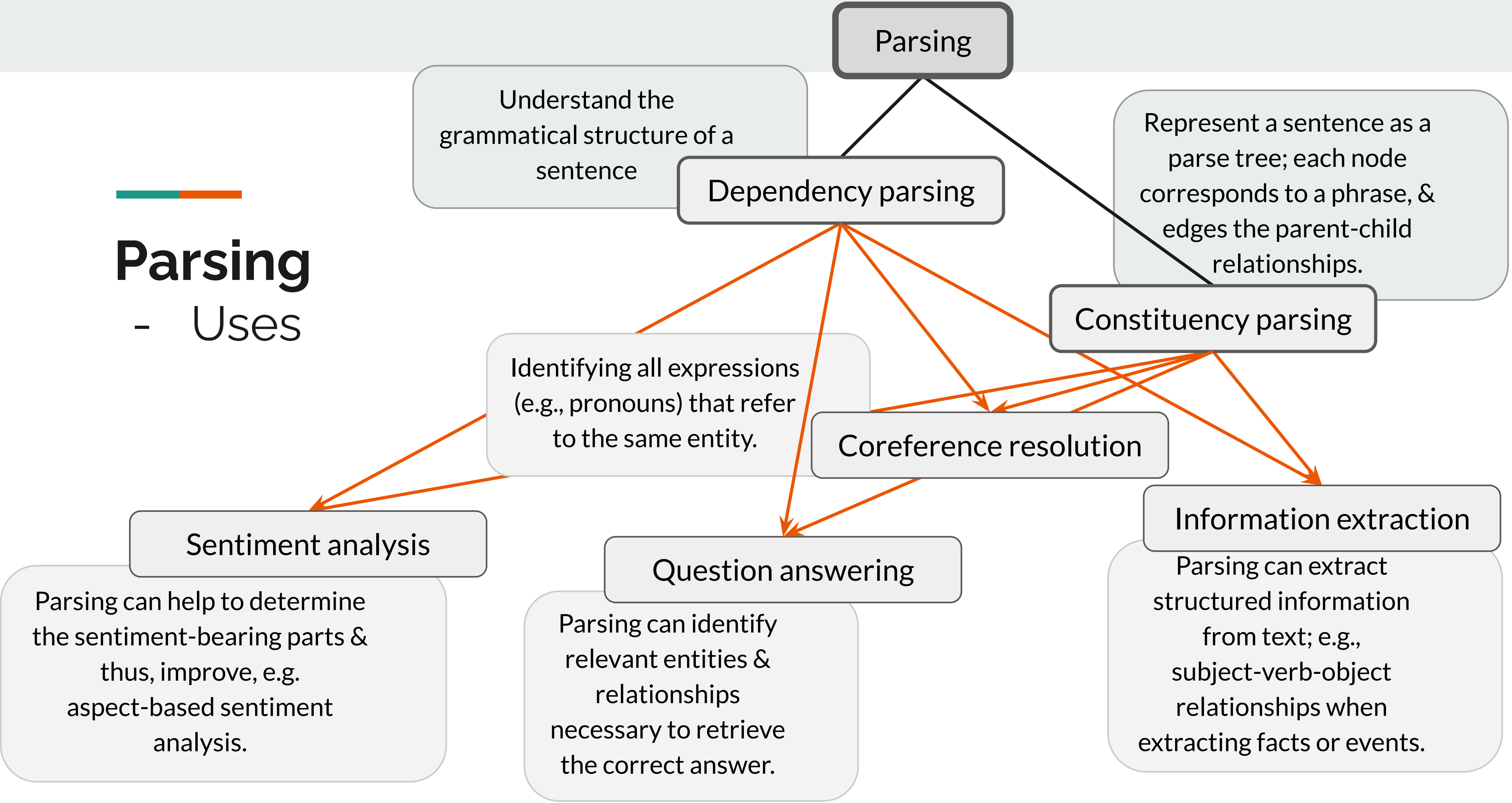
- Describe the syntactic structure of a sentence in terms of the words (or lemmas) in a sentence & associated directed binary grammatical relations among the words.
- Approximates semantic relationships thus making it useful for e.g. *coreference resolution, question answering* & *information extraction.*
- http://nlpprogress.com/english/dependency_parsing.html



No need to know this

NSUBJ	-	Nominal subject
DOBJ	-	Direct object
DET	-	Determinant
NMOD	-	Nominal modifier
CASE	-	Prepositions, postpositions & other case markers

Parsing
- Uses



Part-of-speech Tagging

POS tagging aka grammatical tagging

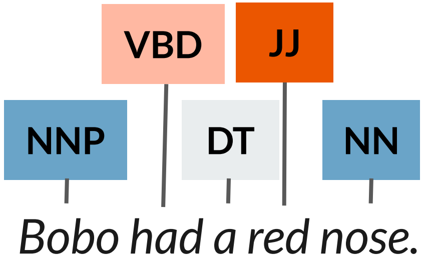
Part-of-speech (POS) aka *word class*, or *syntactic category*.

A POS is a category of words with similar grammatical properties. Common English parts of speech are *noun*, *verb*, *adjective*, *adverb*, *pronoun*, *preposition*, *conjunction*, etc.

POS tagging is a disambiguation task since many words have more than one possible POS.

POS is an important step in many applications. Examples are *lemmatization*, *Named Entity Recognition* (NER), *sentiment analysis*, *question answering*, *word sense disambiguation*, & for grammar & spell checkers.

POS tagging



NNP - proper noun, singular

NN - noun, singular

VBD - verb, past tense

DT - determinant

JJ - adjective

Examples of ambiguous words:
that, *back*, *down*, *put* & *set*.

Ambiguity

- 6 different POS for the word *back*

JJ	VB
earnings growth took a **back** seat	Dave began to **back** toward the door
NN	RP
a small building in the **back**	enable the country to buy **back** about debt
VBD	RB
a clear majority of senators **back** the bill	I was twenty-one **back** then

The Penn Treebank POS tagset

A total of 45 tags
Only a subset in the tables below, see the complete list in Figure 8.2 of *Speech & Language Processing*

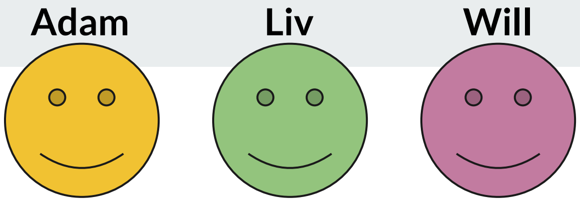
CC	coordinating conjunction	*and*
CD	cardinal number	*1, third*
DT	determiner	*the*
EX	existential there	*there is*
FW	foreign word	*cachaça*
JJ	adjective	*green*
NNP	proper noun, singular	*Hjalmar*
NPS	proper noun, plural	*Vikings*

NN	Noun, singular or mass	*llama*
VB	verb be, base form	*be*
VBD	verb be, past tense	*was, were*
VV	verb, base form	*take*
VVD	verb, past tense	*took*
RB	adverb	*quickly*
RP	particle	*up, off*
MD	modal	*could, will, would*

Two PoS tagging algorithms

For simplicity, we will only use 3 simplified tags:

- **N** noun: e.g. *house, bike, rock, llama, Hjalmar*
  - Notice that we have fused several noun-related tags
- **M** modal verb: e.g. *will, could, would, should, may*
- **V** verb: e.g. *see, run, jump, skip, ride*
  - Notice that we have fused several verb-related tags



The baseline algorithm: most frequent tag

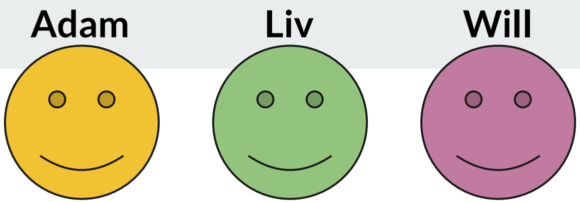
Corpus:

N V N
Liv saw Adam

N V N
Adam saw Will

Tag counts

	N	V
Liv	1	0
Adam	2	0
Will	1	0
saw	0	2



The baseline algorithm: most frequent tag

Corpus:

N V N
Liv saw Adam

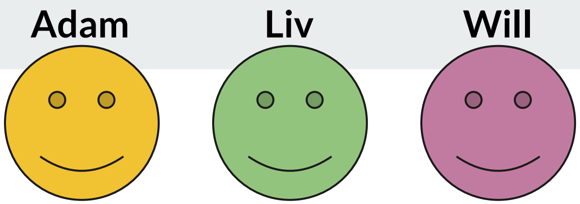
N V N
Adam saw Will

New sentence:

Liv saw Will

Tag counts

	N	V
Liv	1	0
Adam	2	0
Will	1	0
saw	0	2



The baseline algorithm: most frequent tag

Corpus:

N V N
Liv saw Adam

N V N
Adam saw Will

New sentence:

N V N
Liv saw Will

Tag counts

	N	V
Liv	1	0
Adam	2	0
Will	1	0
saw	0	2

The baseline algorithm: most frequent tag

Corpus:

N V N
Liv saw Adam

N V N
Adam saw Will

New sentence:

N V N
Liv saw Will

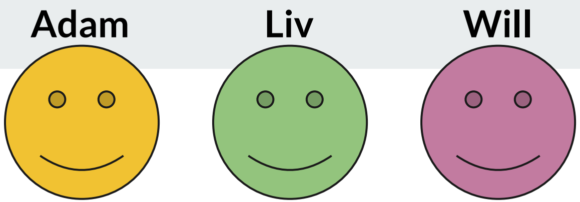
Tag counts

	N	V
Liv	1	0
Adam	2	0
Will	1	0
saw	0	2

This is the baseline algorithm that any new algorithm has to improve upon

- The most-frequent-tag baseline achieves 92.34% accuracy on WSJ, whereas SOTA is around 97%

Question: What is the problem?



A problem with the baseline algorithm

Corpus:

N M V N
Liv will see Adam.

N M V N
Will will see Adam.

N M V N
Adam will see Liv.

New sentence:

Liv will see Will.

Tag counts

	N	M	V
Liv	2	0	0
Adam	3	0	0
Will	1	3	0
see	0	0	3

Question: What context can we use to disambiguate between the tags?

A problem with the baseline algorithm

Document

N M V N
Liv will see Adam.
N M V N
Will will see Adam.
N M V N
Adam will see Liv.

New sentence:

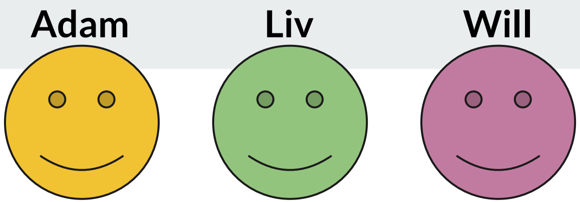
N M V M
Liv will see Will.

Tag counts

	N	M	V
Liv	2	0	0
Adam	3	0	0
Will	1	3	0
see	0	0	3

Will is associated with 2 different tags, i.e. it is ambiguous - **context is needed**

Question: What context can we use to disambiguate between the tags?



Solving this ambiguity is for better tagging algorithms, but not for this lecture.

A problem with the baseline algorithm

Document

N M V N
Liv will see Adam.
N M V N
Will will see Adam.
N M V N
Adam will see Liv.

New sentence:

N M V M
Liv will see Will.

Tag counts

	N	M	V
Liv	2	0	0
Adam	3	0	0
Will	1	3	0
see	0	0	3

Will is associated with 2 different tags, i.e. it is ambiguous - **context is needed**

POS-tagged corpora

The Brown corpus

- A million words taken from 500 texts of different genres.

The WSJ corpus

- A million words taken from Wall Street Journal.

The Switchboard corpus

- 2 million words from telephone conversations.

Tagging was done by an automatic tagger, followed by human corrections.

Dependency Parsing

Wen dependency parsing?

Mainly for improving the performance of downstream tasks, not an application in it self.

With end-to-end models it can be used to enrich the text representation (e.g. marking up text before vectorizing, or as an additional feature vector).

A good parsing model is trained on very large datasets. Thus, a custom model fine-tuned on a small dataset can leverage this important knowledge. Think about it as transferring knowledge from large datasets to small-dataset models (i.e. transfer learning).

Language is recursive

Starting units: words with POS categories.

```
the,  cat,  cuddly,  door,  by
Det   N     Adj      N      P
```

Words combine into phrases with categories.

```
the cuddly cat,    by   the door
Det   Adj   N      P    Det N
       NP          P      NP
```

Phrases can combine into bigger phrases recursively.

```
the cuddly cat by the door
    NP         P   NP
          NP
```

Parse trees

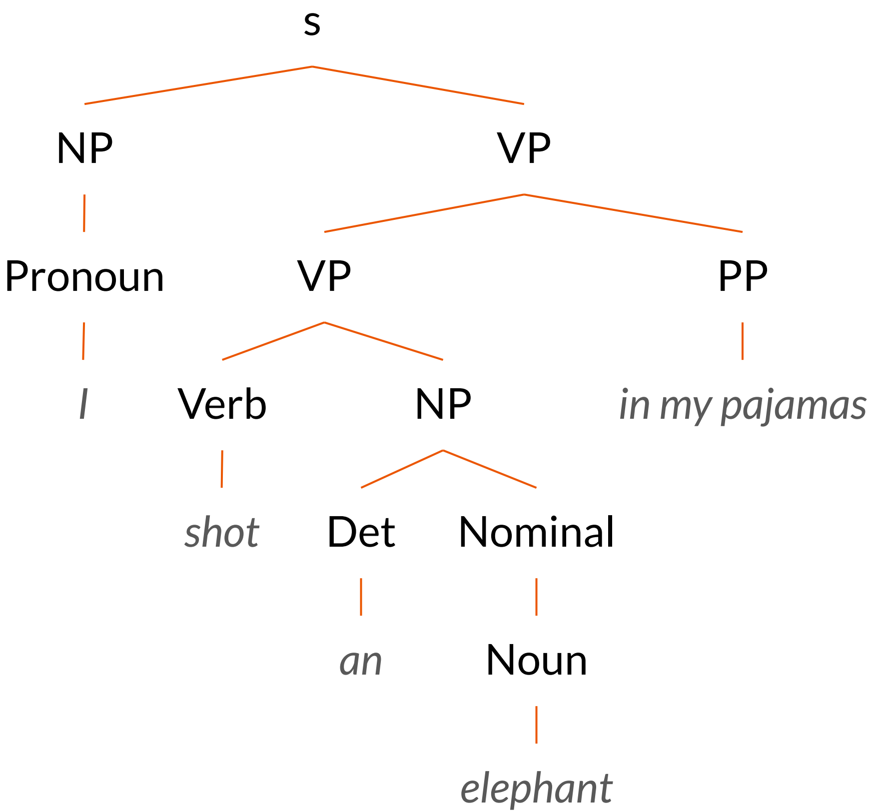
NP - *Noun Phrase*, a sequence of words surrounding a noun.

VP - *Verb Phrase*, a verb & other constituents, e.g. NP & PP.

Det - Determiner (aka article), e.g. **a** stop, **the** flights, etc.

Nominal - follows the determiner & contains pre- & post-noun modifiers. Can consist of only a single noun.

PP - Prepositional Phrase, a preposition followed by an NP.



From Fig 13.2 in Speech & Language Processing (2020)

Shallow parsing

Aka **chunking**, chunks groups of adjacent tokens into phrases on the basis of their POS tags. Well-known chunks are *noun phrases*, *verb phrases*, & *prepositional phrases*.

Noun phrase

- A group of at least 2 words with a noun (e.g. 'she', 'he', 'they', ...) or a pronoun as its head & includes modifiers (e.g. 'the', 'a', 'of them', 'with her').
- A noun phrase plays the role of a noun.
- Examples (modifier noun):

    the man

    a girl

    the doggy in the window

Verb phrase

- A group of at least 2 words consisting of a verb (the head) & arguments that further illustrates the verb tense, action or tone.
- Examples (arguments verb):

    She can smell the pizza

    He has appeared on screen as an actor.

Dependency structure

Dependency structure shows which words ***depend*** on which other words.

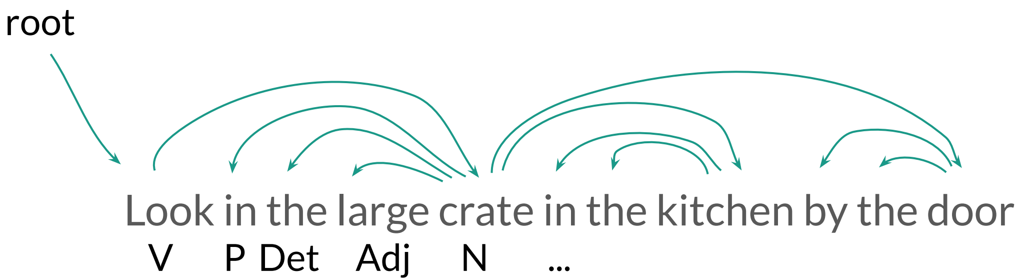
- *Depend*: modify or are arguments of.

Look in the large crate in the kitchen by the door

Dependency structure

Dependency structure shows which words ***depend*** on which other words.

- *Depend*: modify or are arguments of.



Slide credit: Stanford CS224N, Chris Manning

Dependency structure

Dependency structure shows which words ***depend*** on which other words.

- *Depend*: modify or are arguments of.

Look for the large barking dog by the door in a crate

Why do we need dependency parsing?

Why is it important to understand the structure of sentences?

Sentence structure is important to be able to correctly interpret language.

Humans communicate by composing words together into bigger units capable of conveying complex meanings.

Thus, for many NLP tasks it becomes crucial to know what is connected to what, & in what way.

- I.e. what words are arguments & modifiers of other words.

subj - subject
obj - object
nmod - nominal modifier

How it can go wrong

- Prepositional phrase attachment ambiguity

2 alternative meanings.

- The man who was killed had a knife.
- Cops lethally stab a man.

San Jose cops kill man with knife

Text Paper Translate Close

San Jose cops kill man with knife

Ex-college football player, 23, shot 9 times allegedly charged police at fiancee's home

By Hamed Aleaziz and Vivian Ho

A man fatally shot by San Jose police officers while allegedly charging at them with a knife was a 23-year-old former football player at De Anza College in Cupertino who was distraught and depressed, his family said Thursday.

Police officials said two officers opened fire Wednesday afternoon on Phillip Watkins outside his fiancee's home because they feared for their lives. The officers had been drawn to the home, officials said, by a 911 call reporting an armed home invasion that, it turned out, had been made by Watkins himself.

But the mother of Watkins' fiancee, who also lives in the home on the 1300 block of Sherman Street, said she witnessed the shooting and described it as excessive. Faye Buchanan said the confrontation happened shortly after she called a suicide intervention hotline in hopes of getting Watkins medical help.

Watkins' 911 call came in at 5:01 p.m., said Sgt. Heather Randol, a San Jose police spokeswoman. "The caller stated there was a male breaking into his home armed with a knife," Randol said. "The caller also stated he was locked in an upstairs bedroom with his children and requested help from police."

She said Watkins was on the sidewalk in front of the home when two officers got there. He was holding a knife with a 4-inch blade and ran toward the officers in a threatening manner, Randol said.

"Both officers ordered the suspect to stop and drop the knife," Randol said. "The suspect continued to charge the officers with the knife in his hand. Both officers, fearing for their safety and defense of their life, fired at the suspect."

On the police radio, one officer said, "We have a male with a knife. He's walking toward us."

"Shots fired! Shots fired!" an officer said moments later.

A short time later, an officer reported, "Male is down. Knife's still in hand."

Buchanan said she had been prompted to call the

Shoot continues on D8

Back Continue

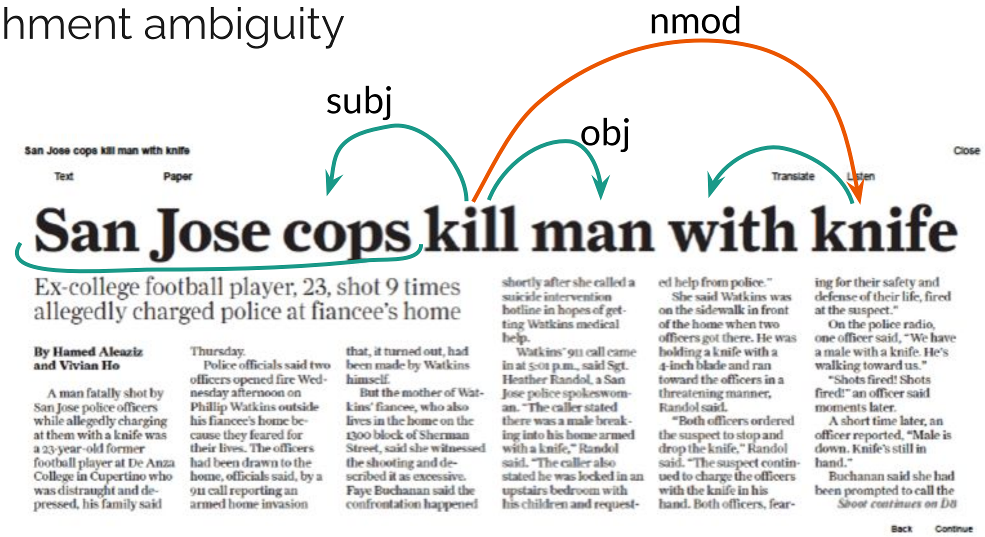
subj - subject
obj - object
nmod - nominal modifier

How it can go wrong

- Prepositional phrase attachment ambiguity

2 alternative meanings.

- The man who was killed had a knife.
- Cops lethally stab a man.



San Jose cops kill man with knife

Text Paper Translate Listen Close

San Jose cops kill man with knife

Ex-college football player, 23, shot 9 times allegedly charged police at fiancee's home

By Hamed Aleaziz and Vivian Ho

A man fatally shot by San Jose police officers while allegedly charging at them with a knife was a 23-year-old former football player at De Anza College in Cupertino who was distraught and depressed, his family said Thursday.

Police officials said two officers opened fire Wednesday afternoon on Phillip Watkins outside his fiancee's home because they feared for their lives. The officers had been drawn to the home, officials said, by a 911 call reporting an armed home invasion that, it turned out, had been made by Watkins himself.

But the mother of Watkins' fiancee, who also lives in the home on the 1300 block of Sherman Street, said she witnessed the shooting and described it as excessive. Faye Buchanan said the confrontation happened shortly after she called a suicide intervention hotline in hopes of getting Watkins medical help.

Watkins' 911 call came in at 5:01 p.m., said Sgt. Heather Randol, a San Jose police spokeswoman. "The caller stated there was a male breaking into his home armed with a knife," Randol said. "The caller also stated he was locked in an upstairs bedroom with his children and requested help from police."

She said Watkins was on the sidewalk in front of the home when two officers got there. He was holding a knife with a 4-inch blade and ran toward the officers in a threatening manner, Randol said.

"Both officers ordered the suspect to stop and drop the knife," Randol said. "The suspect continued to charge the officers with the knife in his hand. Both officers, fearing for their safety and defense of their life, fired at the suspect."

On the police radio, one officer said, "We have a male with a knife. He's walking toward us."

"Shots fired! Shots fired!" an officer said moments later.

A short time later, an officer reported, "Male is down. Knife's still in hand."

Buchanan said she had been prompted to call the

Shoot continues on D8

Back Continue

Question: What are the 2 interpretations of this headline?

How it can go wrong

- Prepositional phrase attachment ambiguity

How it can go wrong

- Prepositional phrase attachment ambiguity

subj obj nmod

Scientists count whales from space.

subj obj nmod

Scientists count whales from space.

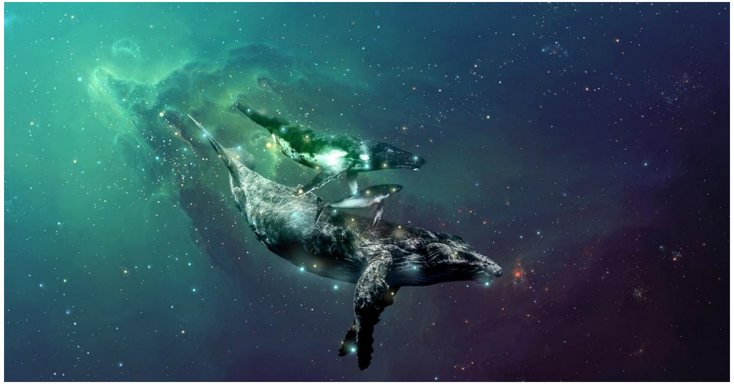


Image credits:
https://www.nrdc.org/onearth/space-whales-look-much-smaller
https://www.laphamsquarterly.org/roundtable/fantastical-allure-space-whale

Example from the Wall Street Journal

How it can go wrong

- Coordination scope ambiguity

Shuttle veteran & longtime NASA executive Fred Gregory appointed to board.

Shuttle veteran & longtime NASA executive Fred Gregory appointed to board.

Question: What are the 2 interpretations of this headline?
Question: What is the cause of the ambiguity?

How it can go wrong

- Coordination scope ambiguity


NATION / WO

PRESIDENT'S FIRST PHYSICAL

Doctor: No heart, cognitive issues

But Trump needs to reduce

6-foot-3 president weighed in at 239 pounds — three pounds heavier than he

with no medical issues." Trump has no heart disease and no family history

How it can go wrong

- Coordination scope ambiguity

&/or

Question: What are the 2 interpretations of this headline?
Question: What is the cause of the ambiguity?

When it goes wrong

- Verb Phrase (VP) attachment ambiguity

Image credit: https://www.theguardian.com/world/2016/jun/29/rio-de-janeiro-mutilated-body-beach-olympics-volleyball

When it goes wrong

- Verb Phrase (VP) attachment ambiguity

Image credit: https://www.theguardian.com/world/2016/jun/29/rio-de-janeiro-mutilated-body-beach-olympics-volleyball

Dependency grammar

Dependency grammar

Dependency grammar assumes that sentence structure consists of relations between words, normally binary asymmetric relations (i.e. arrows) called *dependencies*.

The arrows are often *typed* with the name of grammatical relations (e.g. subject, determiner, prepositional object, etc).

Long history

The first linguist Pāṇini (around 6 - 4th century BCE) creates something like the first dependency grammar.

Written on birch bark!!

Earliest NLP dependency parser is from 1962 (Hays, 1962).

Grammatical relations in dependency parsing

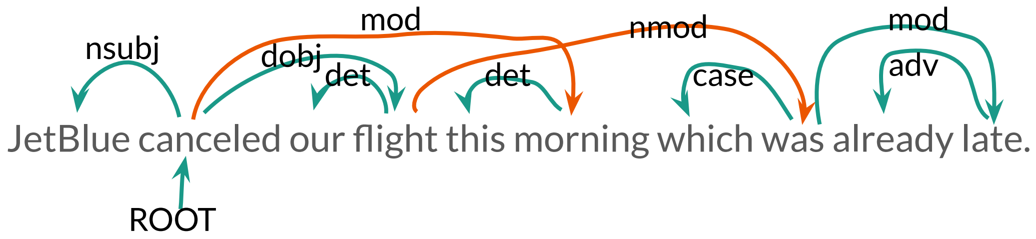
Universal dependency relations
See Universal Dependencies & Stanford typed dependencies manual for a complete list.

Causal argument relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement

Nominal modifier relations	Description
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions & other case markers
Other notable relations	**Description**
CONJ	Conjunction
CC	Coordinating conjunction

Dependency parsing

- A sentence is parsed by deciding for each word what other word (including a fake ROOT) it is a dependent of.
- Constraints are often added
  - Only one word is a dependent of ROOT.
  - No cycles, i.e. A → B & B → A.
- The above constraints makes the dependencies a tree (a directed acyclic graph).
- Sometimes arrows are not allowed to cross (non-projective).



Methods for dependency parsing

Dynamic programming

- Eisner (1996) gives a clever algorithm with complexity $O(n^3)$, by producing parse items with heads at the ends rather than in the middle.

Graph algorithms

- Create a Minimum Spanning Tree (MST) for a sentence.
- McDonald et al.'s (2005) MSTParser scores dependencies independently using an ML classifier.

Constraint Satisfaction

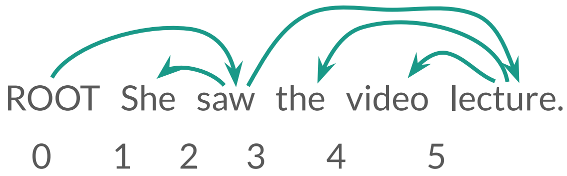
- Edges are eliminated that don't satisfy hard constraints. Karlsson (1990), etc.

Transition-based parsing

- Greedy choice of dependencies guided by ML classifiers, e.g. MaltParser (Nivre et al. 2008). Popular & effective.

Evaluating a dependency parser

Evaluation – UAS



Unlabeled Attachment Score (UAS)

Gold (i.e. ground truth)

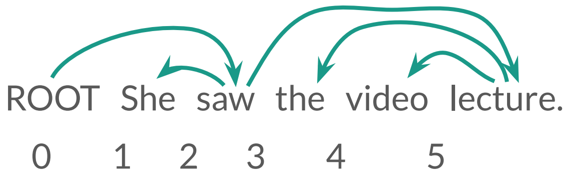
to	from	word	type
1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	obj

Parsed (i.e. predicted)

to	from	word	type
1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

$$Acc = \frac{\#true\ dependencies}{\#dependencies}$$

$$UAS = \frac{4}{5} = 80\%$$



Evaluation – LAS

Labeled Attachment Score (LAS)

Gold (i.e. ground truth)

to	from	word	type
1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	obj

Parsed (i.e. predicted)

to	from	word	type
1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

$$Acc = \frac{\#true\ dependencies}{\#dependencies}$$

$$UAS = \frac{4}{5} = 80\%$$

$$LAS = \frac{2}{5} = 40\%$$

State of the art

Bigger, deeper networks & better hyperparameter tuning.

Beam search.

Self-attention.

Penn Treebank

Models are evaluated on the Stanford Dependency conversion (**v3.3.0**) of the Penn Treebank with **predicted** POS-tags. Punctuation symbols are excluded from the evaluation. Evaluation metrics are unlabeled attachment score (UAS) and labeled attachment score (LAS). UAS does not consider the semantic relation (e.g. Subj) used to label the attachment between the head and the child, while LAS requires a semantic correct label for each attachment.Here, we also mention the predicted POS tagging accuracy.

Model	POS	UAS	LAS	Paper / Source	Code
Label Attention Layer + HPSG + XLNet (Mrini et al., 2019)	97.3	97.42	96.26	Rethinking Self-Attention: Towards Interpretability for Neural Parsing	Official
ACE + fine-tune (Wang et al., 2020)	-	97.20	95.80	Automated Concatenation of Embeddings for Structured Prediction	Official
HPSG Parser (Joint) + XLNet (Zhou et al, 2020)	97.3	97.20	95.72	Head-Driven Phrase Structure Grammar Parsing on Penn Treebank	Official
Second-Order MFVI + BERT (Wang et al., 2020)	-	96.91	95.34	Second-Order Neural Dependency Parsing with Message Passing and End-to-End Training	Official
CVT + Multi-Task (Clark et al., 2018)	97.74	96.61	95.02	Semi-Supervised Sequence Modeling with Cross-View Training	Official
CRF Parser (Zhang et al., 2020)	-	96.14	94.49	Efficient Second-Order TreeCRF for Neural Dependency Parsing	Official
Second-Order MFVI (Wang et al., 2020)	-	96.12	94.47	Second-Order Neural Dependency Parsing with Message Passing and End-to-End Training	Official
Left-to-Right Pointer Network (Fernández-González and Gómez-Rodríguez, 2019)	97.3	96.04	94.43	Left-to-Right Dependency Parsing with Pointer Networks	Official

Image credit: https://github.com/sebastianruder/NLP-progress/blob/master/english/dependency_parsing.md

More linguistic Features

Enriched features

Raw text presents challenges for NLP: rare words, varying word forms with similar meanings, & order-dependent interpretations. Enriching text with linguistic features can improve BOW models. This involves incorporating information about word function & relationships, such as POS tags & dependency relations, to enhance the model.

Word shapes

Problem

- Many words are rare.

Solution

- Featurize words by mapping them to simpler representations that captures word-attributes like length, capitalization, numerals, internal punctuation, Greek letters, etc.

Word	**shape**
Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

CPA1 & Varicella-zoster.

Word shapes

Featurize words by mapping them to simpler representations that captures word-attributes like length, capitalization, numerals, internal punctuation, Greek letters, etc.

First & last 2 characters are most important. The middle of words are replaced with unique codes in a specific order.

A, B, C, ..., Z → X
a, b, c, ..., z → x
0, 1, 2, ... → d
-, :, ., ... → -, :, ., ...

Word	shape
Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

CPA1 & Varicella-zoster.

Word substrings

Examples

- `oxa` – great precision for **drug** names, e.g. *Cotrimoxazole.*
- `:` – good precision for **movie** names, e.g. *Alien Fury: Countdown to Invasion.*
- `field` – ok precision for **person** names, e.g. *Wethersfield* & *Mansfield* & Hadfield.

Image credit:
Robert Markowitz - This image was catalogued by Johnson Space Center of the United States National Aeronautics & Space Administration (NASA) under Photo ID: JSC2011-E-203354., Public Domain, https://commons.wikimedia.org/w/index.php?curid=19962522

Tutorial

`MMAI5400_class06_enrichFeats.ipynb`