Chapter NLP:V

V. Words

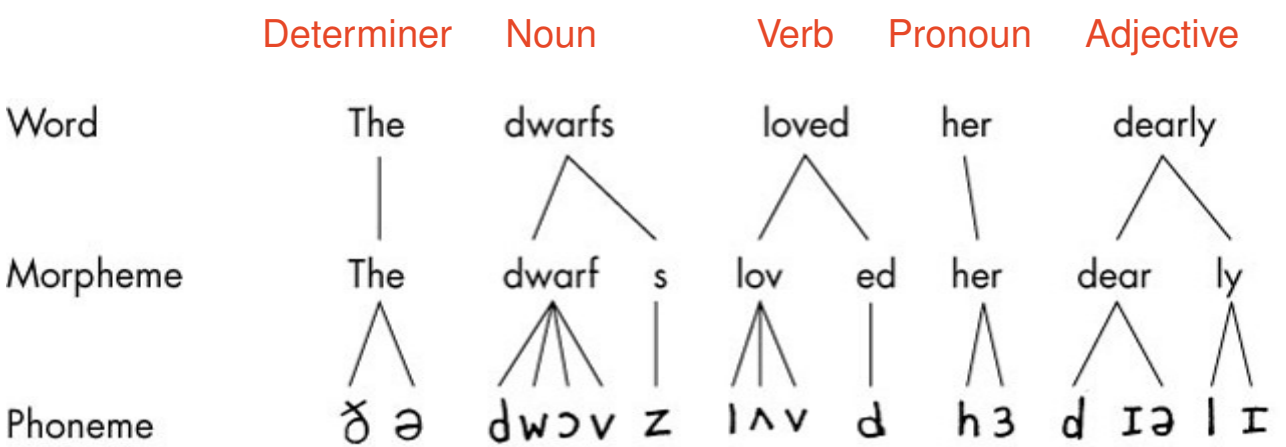
- Morphology
- Word Classes
- Named Entities

Word Classes

Definition

A word class is a set of lexical items with similar formal (grammatical) properties.

- Also called part of speech, grammatical category, lexical category, or syntactic category. roughly synonymous
- Common properties are morphology and semantic or syntactic behavior.
- Often serve an encoding of the formal properties of words for programs.



Word Classes

Traditional grammar

The traditional English grammar lists 9 word classes split into two groups: function or content classes.

1. Content classes: or open/form/lexical
   - Accept new members, can hold infinitely many items per class.
   - Nouns, verbs, adjectives, adverbs.
2. Function classes: or closed/structure
   - Number of members is fixed.
   - Prepositions, pronouns, determiners, conjunctions and interjections.
   - As language evolves, changes may also happen in closed classes.

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN					
UD					

Word Classes

Traditional grammar: Example

Word Class: Noun
Definition: A noun is a word used for naming some person or thing. [Cambridge]
Examples: Paris, man, house, height.

- The definition is incomplete: What about places? What about abstract qualities (beauty) and actions (a thump)?
- No reference is made to morphology or syntax.
- The class does not differentiate (grammatical) properties: Plural nouns and proper nouns are all just **noun**.

Remarks:

- An **adjective** is a word used to qualify a noun [...] to restrict the application of a noun by adding something to its meaning. Examples: `fine, brave, three, the`.
  - The definition is vague and allows many elements with different grammatical properties (`the, my, all`), and even nouns in certain constructions (`her brother the butcher`).
  - No reference is made to morphology or syntax.
- A **verb** is a word used for saying something about some person or thing. Examples: `make, know, buy, sleep`.
  - On this definition, there is little difference between a verb and an adjective. Some grammars prefer to talk about 'doing words' or 'action words', but this seems to exclude the many state verbs, such as `know, remember, be`.
  - No reference is made to morphology or syntax.
- An **adverb** is a word used to qualify any part of speech except a noun or pronoun. Examples: `today, often, slowly, very`.
  - Adverbs are often said to qualify (or 'modify') verbs – which is inadequate for such words as `very` and `however`.
  - This definition hardly applies to interjections or examples like `the very man` and `slovenly me`.
  - No reference is made to morphology or syntax.

Remarks:

- A **pronoun** is a word used instead of a noun or noun-equivalent (i.e. a word which is acting as a noun). Examples: `this, who, mine`.
  - Pronouns are used instead of noun phrases, not just nouns. `He` refers to the whole of the phrase `the big lion`, not just the word `lion`, since we cannot say `the big he`.
  - No reference is made to morphology or syntax.
- A **preposition** is a word placed before a noun or noun-equivalent to show in what relation the person or thing stands to something else. Examples: `on, to, about, beyond`.
  - This gives a clear syntactic criterion. However, prepositions go before noun phrases, rather than nouns, and may also be used in other parts of the sentence. As with nouns, more than just persons and things are involved.
- A **conjunction** is a word used to join words or phrases together, or one clause to another clause, Examples: `and, before, as well as`.
  - Prepositions also have a joining function (`the man in the garden`).
  - Conjunction Junction, what's your function?
- An **interjection** is a word or sound thrown into a sentence to express some feeling of the mind. Examples: `Oh!, Bravo!, Fie!`.
  - Interjections do not enter into the construction of sentences. Despite the emotional function of these words, they still need to be considered as part of sentence classification.

Word Classes

Tagsets

The classes of the traditional grammar are not suited for language processing.

- Several advanced class schemes (called tagsets) exist.
- They distinguish between 17 up to 100+ word classes.
  - Penn Treebank tagset 36 tags
  - Universal POS tags 17 tags
  - CLAWS tagsets CLAWS1: 132, CLAWS2: 166, C5: 60, C6: 160, C8, ...
  - Brown Corpus 87 tags
  - Penn Treebank II 41 tags
  - British National Corpus 61 tags
  - BNC Sampler 146 tags
- Corpora for part of speech tagging are manually annotated, for example:
  - The 1 million word Brown corpus in the 1960s.
  - The 100 million word British National Corpus.

Word Classes

Penn Treebank tagset [upenn]

Idea: Assign a tag to each combination of class of the traditional grammar and their observed grammatical properties.

NN	Noun in singular
NNS	Noun in plural
NNP	Proper noun
NNPS	Proper noun in plural

Word Classes

Penn Treebank tagset [upenn]

Idea: Assign a tag to each combination of class of the traditional grammar and their observed grammatical properties for all classes.

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
WRB	Wh-adverb

TO	to (Preposition)
RP	Particle
POS	Possessive ending
MD	Modal
PRP	Personal pronoun
PRP$	Possessive pronoun
WP	Wh-pronoun
WP$	Possessive wh-pronoun
DT	Determiner
PDT	Predeterminer
WDT	Wh-determiner
CC	Coordinating conjunction
IN	Preposition or subordinating conjunction
UH	Interjection

Word Classes

Penn Treebank tagset [upenn]

Idea: Assign a tag to each combination of class of the traditional grammar and their observed grammatical properties for all classes. Add classes for everything else we find while annotating text.

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
WRB	Wh-adverb

TO	to
RP	Particle
POS	Possessive ending
MD	Modal
PRP	Personal pronoun
PRP$	Possessive pronoun
WP	Wh-pronoun
WP$	Possessive wh-pronoun
DT	Determiner
PDT	Predeterminer
WDT	Wh-determiner
CC	Coordinating conjunction
IN	Preposition or subordinating conjunction
UH	Interjection
CD	Cardinal number
EX	Existential there
FW	Foreign word
LS	List item marker
SYM	Symbol

Word Classes

Penn Treebank tagset [upenn]

The Penn tagset covers more grammatical properties and is still frequently used.

The Penn tagset is also English-centric:

- What does it mean and why is this a problem?

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN	DT	NNS	VBD	PRP	RB
UD					

Word Classes

Penn Treebank tagset [upenn]

The Penn tagset covers more grammatical properties and is still frequently used.

The Penn tagset is also English-centric:

- With this strategy, each language needs it's own tagset.
- A different tagger needs to be developed for each language.
  Very difficult for low-resource languages, see electronic colonialism.
- It violates Chomsky's theory of universal grammar, which is very beloved by computer scientists.

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN	DT	NNS	VBD	PRP	RB
UD					

Word Classes

Universal Dependencies tagset [UD]

Idea: Seperate form from function: Use only few, universal POS tags and many lexical and grammatical properties which can be freely assigned to any tag.

Part-of-speech Tag	
NOUN	Noun
PROPN	Proper noun
VERB	Verb
AUX	Auxiliary
ADJ	Adjective
ADV	Adverb
ADP	Adposition (Preposition)
PART	Particle
PRON	Pronoun
DET	Determiner
CCONJ	Coordinating conjunction
SCONJ	Subordinating conjunction
INTJ	Interjection
NUM	Numeral
PUNCT	Punctuation
SYM	Symbol
X	Other

Lexical Features
PronType
NumType
Poss
Reflex
Foreign
Abbr
Typo

Inflextional Features	
Nominal	Verbal
Gender	VerbForm
Animacy	Mood
NounClass	Tense
Number	Aspect
Case	Voice
Definite	Evident
Degree	Polarity
	Person
	Polite
	Clusivity

Which of these do not exist in English?

Word Classes

Universal Dependencies tagset [UD]

Idea: Seperate form from function: Use only few, universal POS tags and many lexical and grammatical properties which can be freely assigned to any tag.

	The	dwarfs	loved	her	dearly
TG	determiner	noun	verb	pronoun	adverb
PENN	DT	NNS	VBD	PRP	RB
UD	DET; PronType: Art	NOUN; Number: Plur	VERB; Tense: Past	PRON	ADV

Word Classes

Ambiguities

- About 85% of the vocubulary (types) belong to only one word class.
- The others are ambiguous.

The back door	→	adjective, JJ
On my back	→	noun, NN
Win the voters back	→	adverb, RB
Said to back the bill	→	verb, VB

- However, ambiguous types are much more frequent. About half of the tokens are ambiguous.

Remarks:

- About the termininology, the Cambridge Encyclopedia of the English Language (CUP) writes:

  "When linguists began to look closely at English grammatical structure in the 1940s and 1950s, they encountered so many problems of identification and definition that the term part of speech soon fell out of favor, word class being introduced instead. Word classes are equivalent to parts of speech, but defined according to strict linguistic criteria." – (Crystal (2003)

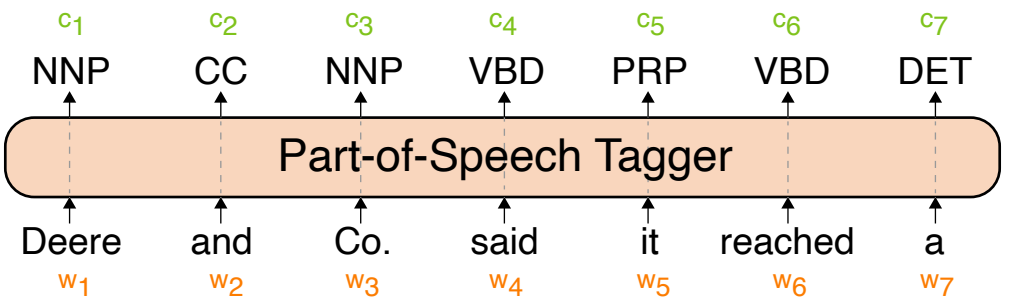
- "There is no single correct way of analyzing words into word classes. . . Grammarians disagree about the boundaries between the word classes, and it is not always clear whether to lump subcategories together or to split them. For example, in some grammars pronouns are classed as nouns, whereas in other frameworks they are treated as a separate word class." – Aarts, Chalker, Weiner (2014) The Oxford Dictionary of English Grammar. OUP
- The 9 word classes of the traditional grammar are usually attributed to *Dionysius Thrax of Alexandria (100 B.C.)*, who wrote in his *Techne* on Greek about eight parts of speech: noun, verb, pronoun, preposition, adverb, conjunction, participle, and article.

Word Classes

Part-of-Speech Tagging

A tagger is a program that assigns to each token w_i from an input sequence $w_1, \ldots, w_n$ a tag c_i from a tagset T.

- The output is a sequence $c_1, \ldots, c_n$ with the same length as the input.



Word Classes

Part-of-Speech Tagging

A tagger is a program that assigns to each token w_i from an input sequence $w_1, \ldots, w_n$ a tag c_i from a tagset T.

- The output is a sequence $c_1, \ldots, c_n$ with the same length as the input.

Part-of-speech tags can inform us about syntax and semantics of a sequence:

- the intended sense of a word

  `apple` (single noun, NN) **vs.** `Apple` (proper noun, NNP)
- the applied morphemes (lemmatization)

  `sigh` (verb base form, VB) **vs.** `sighed` (verb past tense *or* past participle, VBD or VBN)
- the meaning of a sentence (shallow parsing)
- the correct pronunciation (speech synthesis)
  - OBject vs. obJECT, CONtent vs. conTENT

Word Classes

Part-of-Speech Tagging: Maximum Likelihood Estimate

Idea: Most types are unambiguous. Most ambiguous types have one very likely tag. So we can do a Maximum Likelihood Estimate (MLE):

- Tag each token w_i with the word class c_i it appears in most often:

$$c_i = \operatorname*{argmax}_{c_j \in T} \frac{\text{count}(c_j, w_i)}{\text{count}(w_i)}$$

- Unknown words are often tagged as proper nouns.
- This is sometimes called the most frequent class baseline.
- This baseline has 92% accuracy on UD over 15 languages [Wu and Dredze 2019].
- Humans reach ca. 97% accuracy.

Word Classes

Part-of-Speech Tagging: Brill Tagger [Brill 1992]

Idea: Improve MLE and apply rules to correct errors in the ambiguous cases.

The Brill Tagger is an "error-driven transformation-based tagger". It iteratively applies rules $\texttt{c}_i$ $\texttt{c}_j$ `<Premise>`.

1. Initially tag a sequence with MLE.
2. If a token is tagged with $\texttt{c}_i$ and fulfills the `<Premise>`, replace $\texttt{c}_i$ with $\texttt{c}_j$.
3. Repeatedly iterate the sequence and apply the rules in order until the stopping criterion is reached.

The rules are learned from errors made on a pre-tagged corpus.

- If a false tag $\texttt{c}_i$ is encounterd in the pre-tagged corpus, create several rules with the correct tag $\texttt{c}_j$ and a fulfilled `<Premise>`.
- Correct a test corpus with each new rule.
- If a rule does not increase accuracy, discard it.

Word Classes

Part-of-Speech Tagging: Brill Tagger [Brill 1992]

The premises are manually created from templates.

Premise templates:

`context x`	A word in `context` is tagged `x`.
`property`	The word has a certain `property`.
`context property`	A word in `context` has a certain `property`.
`context`	One or any of $i \in [1, 3]$ preceding or following word(s).
`property TRUE|FALSE`	Capitalized word.

Example rules:

```
TO   IN   next-tag AT                  NN   VB   prev-tag TO
VBN  VBD  prev-word-is-cap TRUE        TO   IN   next-word-is-cap TRUE
VBD  VBN  prev-1-or-2-or-3-tag HVD     NN   VB   prev-tag MD
VB   NN   prev-1-or-2-tag AT
```

Problem: How to handle unknown (previously unseen) words?

Word Classes

Part-of-Speech Tagging: Brill Tagger [Brill 1994]

Idea: Create rules to tag unknown tokens. c_i may be `UNK` for unknown.

Premise templates for unknown tokens:

`affix x constraint`	Token fulfills `constraint` regarding affix of at most 4 chars.
`context word`	A word appears in `context`.
`char x`	Character `x` occurs in word.
`constraint`	When deleting or adding affix `x`, word found in dictionary. Else, affix `x` occurs in token.

Example rules:

```
NN    CD    char .
NN    JJ    char -
UNK   ADJ   suffix -ly addition
UNK   RB    suffix -ly occurs
```

```
NN   NNS   suffix -s occurs
NN   VBN   suffix -ed occurs
NN   VBG   suffix -in occurs
```

Word Classes

Part-of-Speech Tagging: Brill Tagger [Brill 1994]

Idea: Create rules to tag unknown tokens. c_i may be `UNK` for unknown.

Premise templates for unknown tokens:

`affix x constraint`	Token fulfills `constraint` regarding affix of at most 4 chars.
`context word`	A word appears in `context`.
`char x`	Character `x` occurs in word.
`constraint`	When deleting or adding affix `x`, word found in dictionary. Else, affix `x` occurs in token.

Example rules:

```
NN   CD   char .                    NN   NNS   suffix -s occurs
NN   JJ   char -                    NN   VBN   suffix -ed occurs
UNK  ADJ  suffix -ly addition       NN   VBG   suffix -in occurs
UNK  RB   suffix -ly occurs
```

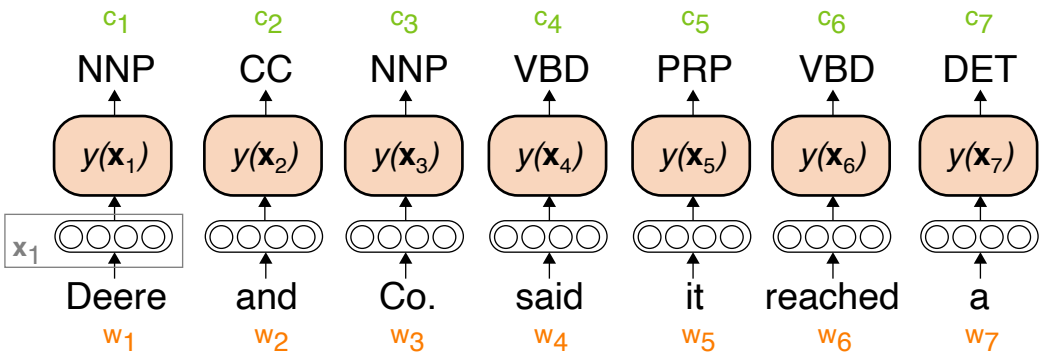
Brill has the typical pitfalls of rule-based systems.

- Poorly generalizes to unknown, new, or misspelled words.
- Has to be (expensively) crafted for every language.

Word Classes

Part-of-Speech Tagging: Token Classification

Idea: Determine the tag c_i of token w_i via classification. Classify $c_i = y(\mathbf{x}_i)$ based on a representation $\mathbf{x}_i$ of w_i:



- $y(\mathbf{x}_i)$ could simply be a maximum likelihood estimator. Works well for POS. Why?
- $\mathbf{x}_i$ can be the vocabulary index of w_i, a contextualized word vector, ...

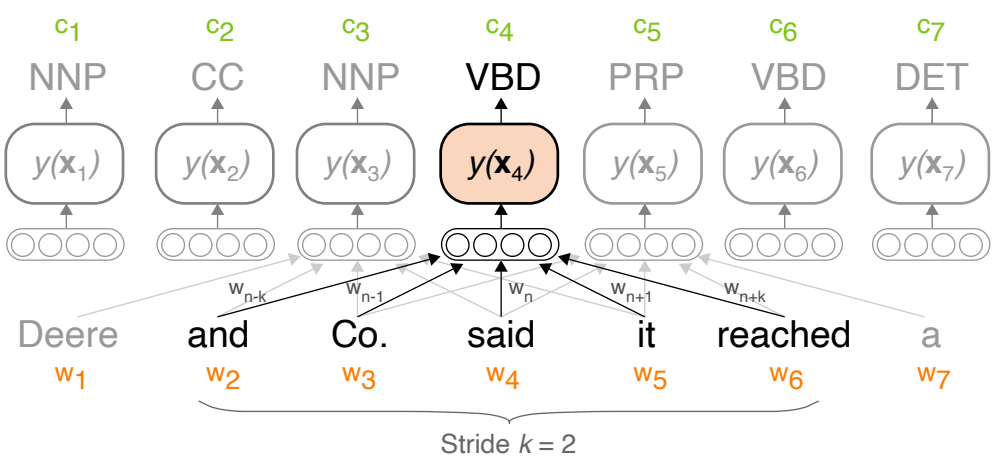
Word Classes

Part-of-Speech Tagging: Token Classification

Idea: Classify $c_i = y(\mathbf{x}_i)$ based on a representation $\mathbf{x}_i$ of the span $w_{i-k}, \ldots, w_{i+k}$ with stride k.

- $\mathbf{x_i}$ is the feature vector for w_i.
- Features can be adapted and extended from Brill's premise templates:

$\mathbf{x}_i^1$	Index of w_i	$\mathbf{x}_i^6$	w_i is in vocabulary	$\mathbf{x}_i^{11}$	w_i is in some word list
$\mathbf{x}_i^2$	Index of w_{i-1}	$\mathbf{x}_i^7$	w_i contains .	$\mathbf{x}_i^{12}$	w_{i-1} has suffix `-ly`
$\mathbf{x}_i^3$	Index of w_{i+1}	$\mathbf{x}_i^8$	w_i is capitalized	$\mathbf{x}_i^{13}$	w_{i-1} w/o `-ly` is in vocabulary
$\mathbf{x}_i^4$	Index of w_{i-2}	$\mathbf{x}_i^9$	tag of w_{i-1}	$\mathbf{x}_i^{14}$	w_{i-1} with `-ly` is in vocabulary
$\mathbf{x}_i^5$	Index of w_{i+2}	$\mathbf{x}_i^{10}$	tag of w_{i-2}	$\mathbf{x}_i^{15}$	...



Word Classes

Token Classification

Idea: Classify $c_i = y(\mathbf{x}_i)$ based on a representation $\mathbf{x}_i$ of the span $w_{i-k}, \ldots, w_{i+k}$ with stride k.

Typical features in $\mathbf{x}_i$ are :

1. For $w_{i-k}, \ldots, w_{i+k}$ the vocabulary indices, pre- and suffixes, capitalization, or occurances in word lists.
2. For $w_{i-k}, \ldots, w_{i-1}$ the already determined tags.

Tagging with span classification:

1. Pad all sequence on both sides with a special token.
   `[PAD] [PAD]` Dere and Co. said . . . `[PAD] [PAD]`
2. Model the training examples $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\}$ for all sequences in training dataset.
3. Train the classifier $y(\mathbf{x})$.
4. Predict c_i identically by padding and then iterating over the words.

Remarks:

- Note the notation: In general sequence processing (like machine translation or language modelling), w is used for the input and y for the output. In tagging, the output is often denominated with l. We use c in the examples to be consistent to machine learning notation: c denotes classes, $\mathbf{x}$ feature vectors, and y the classifier.
- Span-based token classification is also often used for transition-based syntax parsers like arc-standard.
- Common special tokens are `[PAD]`, `[UNK]`, `[MASK]`, `[SEP]`, and `[EOS]`. These are usually treated as individual words by the tokenizer, they are not preprocessed, and they have their own vocabulary indices.
- Its possible to offset the stride to include more left-side than right-side context, or vice versa.
- The classifier $y(\mathbf{x})$ can also be a set of hand-crafted rules.

Word Classes

Part-of-Speech Tagging

Original text

A relevant document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales.

Tagged text with Brill tagger

A/**DT** relevant/**JJ** document/**NN** will/**MD** describe/**VB** marketing/**NN** strategies/**NNS** carried/**VBD** out/**IN** by/**IN** U.S./**NNP** companies/**NNS** for/**IN** their/**PRP$** agricultural/**JJ** chemicals/**NNS** ,/, report/**NN** predictions/**NNS** for/**IN** market/**NN** share/**NN** of/**IN** such/**JJ** chemicals/**NNS** ,/, or/**CC** report/**NN** market/**NN** statistics/**NNS** for/**IN** agrochemicals/**NNS** ,/, pesticide/**NN** ,/, herbicide/**NN** ,/, fungicide/**NN** ,/, insecticide/**NN** ,/, fertilizer/**NN** ,/, predicted/**VBN** sales/**NNS** ,/, market/**NN** share/**NN** ,/, stimulate/**VB** demand/**NN** ,/, price/**NN** cut/**NN** ,/, volume/**NN** of/**IN** sales/**NNS** ./.

Remarks:

- Part-of-speech tagging can be solved with any generic sequence labeling model, like Hidden Markov Models, CRFs, RNNs, and Transformer models.
- POS-Tagging can also be interpreted as conditional language modelling. Intuition: The sentence and tags are the same sentence but in different languages, so we encode the sentence, pass it as condition to a language model that speaks “Part-of-Speech”, and let it generate the tag sequence.
- The Brill Tagger can be seen as a special case of a sequence classifier, where the premises are the features and the order of the rules are the weights.
- The state of the art in part of speech tagging can be reviewed at aclweb.org, paperswithcode, or NLPprogress. Most taggers reported are based on statistical sequence models rather than rules. However, many taggers proposed are not included, including the Brill tagger.