Language Technology
http://cs.lth.se/edan20/
Chapter 10: Techniques for Sequence Prediction

Pierre Nugues

Pierre.Nugues@cs.lth.se
http://cs.lth.se/pierre_nugues/

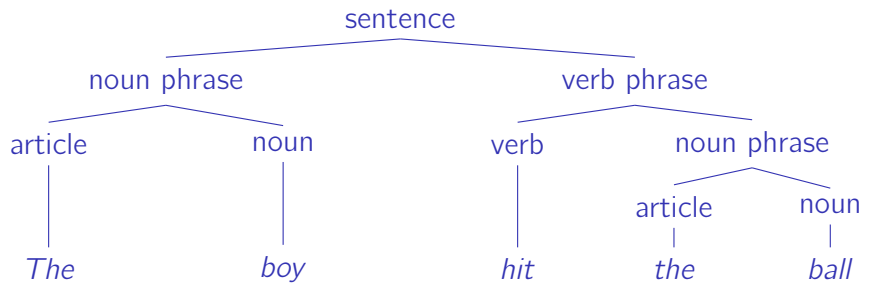
September 22, 2022

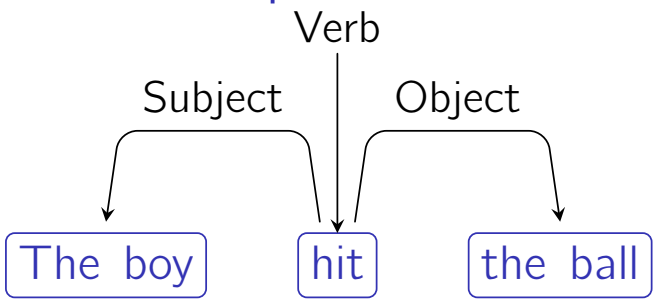
Sequence Prediction

- In this section, we will see how we can apply sequence prediction for part-of-speech tagging to partial parsing.
- We will extend this to other another task we already saw: tokenization

Partial Parsing

- Parsing is the analysis of the relation between the words of a sentence using constituents or dependencies.
- For the sentence *The boy hit the ball*, this corresponds to:



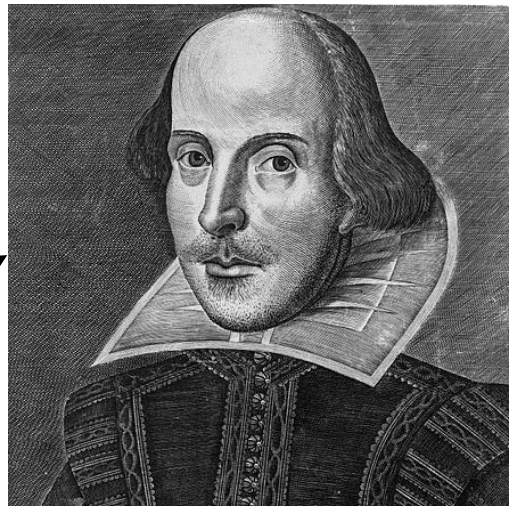


- Parsing might be difficult and useless
- The analysis of parts of a sentence may be enough for many tasks

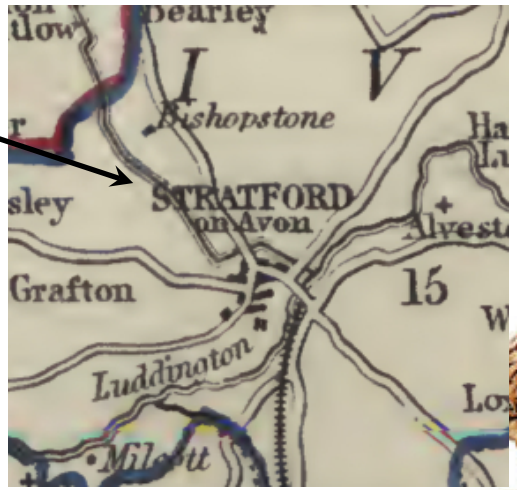
Multiwords

Type	English	French
Prepositions Adverbs Conjunctions	*to the left hand side* *because of*	*À gauche de* *à cause de*
Names	*British gas plc.*	*Compagnie générale d'électricité SA*
Titles	*Mr. Smith* *The President of the United States*	*M. Dupont* *Le président de la République*
Verbs	*give up* *go off*	*faire part* *rendre visite*

Named Entities: Proper Nouns



William Shakespeare was born and brought up in Stratford-upon-Avon



Others Entities: Common Nouns

Meeting with our guest on the landing at lunchtime

Multiword Annotation

The Message Understanding Conferences (MUC), a benchmarking competition organized by the US military, defined an annotation scheme.
The MUC annotation restricts the annotation to information useful to the funding source: names (named entities), time expressions, and money quantities.
The annotation scheme defines an XML element for three classes: `<ENAMEX>`, `<TIMEX>`, and `<NUMEX>` with which it brackets the relevant phrases in a text.
The phrases can be real multiwords, consisting of two or more words, or restricted to a single word.

<ENAMEX>

The <ENAMEX> element identifies proper nouns and uses a TYPE attribute with three values to categorize them: ORGANIZATION, PERSON, and LOCATION as in

- The <ENAMEX TYPE="PERSON">Clinton</ENAMEX> government
- <ENAMEX TYPE="ORGANIZATION">Bridgestone Sports Co.</ENAMEX>
- <ENAMEX TYPE="ORGANIZATION">European Community</ENAMEX>
- <ENAMEX TYPE="ORGANIZATION">University of California</ENAMEX> in <ENAMEX TYPE="LOCATION">Los Angeles</ENAMEX>

Named Entities

The detection of named entities and multiwords with regular expressions is an extension of word spotting.
Just as for word spotting, we store them in a Python dictionary.
To get a list of names, we can use geographical or name dictionaries, called **gazetteers**.
We can also model patterns, for example for:

```
'<ENAMEX> M. Dupont </ENAMEX>'
```

and

```
'<NUMEX> 200 euros </NUMEX>'
```

Noun Groups

English	French	German
The waiter *is bringing* **the very big dish** *on* **the table**	**Le serveur** *apporte* **le très grand plat** *sur* **la table**	**Der Ober** *bringt* **die sehr große Speise** *an* **den Tisch**
Charlotte *has eaten* **the meal** *of* **the day**	**Charlotte** *a mangé* **le plat** *du* **jour**	**Charlotte** *hat* **die Tagesspeise** *gegessen*

Verb Groups

English	French	German
The waiter **is bringing** *the very big dish on the table*	*Le serveur* **apporte** *le très grand plat sur la table*	*Der Ober* **bringt** *die sehr große Speise an den Tisch*
Charlotte **has eaten** *the meal of the day*	*Charlotte* **a mangé** *le plat du jour*	*Charlotte* **hat** *die Tagesspeise* **gegessen**

Segment Recognition

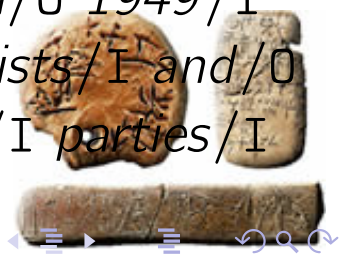
Group detection – chunking –:

Brackets: [$_{NG}$ The government $_{NG}$] has [$_{NG}$ other agencies and instruments $_{NG}$] for pursuing [$_{NG}$ these other objectives $_{NG}$] .

Tags: *The*/I *government*/I *has*/O *other*/I *agencies*/I *and*/I *instruments*/I *for*/O *pursuing*/O *these*/I *other*/I *objectives*/I ./O

Brackets: Even [$_{NG}$ Mao Tse-tung $_{NG}$] [$_{NG}$'s China $_{NG}$] began in [$_{NG}$ 1949 $_{NG}$] with [$_{NG}$ a partnership $_{NG}$] between [$_{NG}$ the communists $_{NG}$] and [$_{NG}$ a number $_{NG}$] of [$_{NG}$ smaller, non-communists parties $_{NG}$] .

Tags: *Even*/O *Mao*/I *Tse-tung*/I *'s*/B *China*/I *began*/O *in*/O *1949*/I *with*/O *a*/I *partnership*/I *between*/O *the*/I *communists*/I *and*/O *a*/I *number*/I *of*/O *smaller*/I *,*/I *non-communists*/I *parties*/I ./O



Other Chunking Schemes

Tjong and Venstra (1999) created 3 other schemes: IOB1, IOB2, IOE1, and IOB2. A 5th tagset, BIOES, is gaining popularity:

- IOB1 : Inside, Outside, Between
- IOB2 : Begin, Inside, Outside, possibly the most popular
- IOE1 : Inside, Outside, End (between two chunks)
- IOE2 : Inside, Outside, End
- BIOES : Begin, Inside, Outside, End, and Singleton, the most efficient one.

Other Chunking Schemes

IOB1 Even/O Mao/I Tse-tung/I 's/B China/I began/O in/O 1949/I with/O a/I partnership/I between/O the/I communists/I and/O a/I number/I of/O smaller/I, non-communists/I parties/I

IOB2 Even/O Mao/B Tse-tung/I 's/B China/I began/O in/O 1949/B with/O a/B partnership/I between/O the/B communists/I and/O a/B number/I of/O smaller/B, non-communists/I parties/I

IOE1 Even/O Mao/I Tse-tung/E 's/I China/I began/O in/O 1 949/I with/O a/I partnership/I between/O the/I communists/I and/O a/I number/I of/O smaller/I, non-communists/I parties/I

BIOES Even/O Mao/B Tse-tung/E 's/B China/E began/O in/O 1949/S with/O a/B partnership/E between/O the/B communists/E and/O a/B number/E of/O smaller/B, non-communists/I parties/E

IOB Annotation for Named Entities

CoNLL 2002	
Words	**Named entities**
Wolff	B-PER
,	O
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

CoNLL 2003			
Words	**POS**	**Groups**	**Named entities**
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Segment Categorization

Tags extendible to any type of chunks: nominal, verbal, etc.
For the IOB scheme, this means tags such as I.Type, O.Type, and B.Type, Types being NG, VG, PG, etc.
In CoNLL 2000, ten types of chunks

Word	POS	Group	Word	POS	Group
He	PRP	B-NP	*to*	TO	B-PP
reckons	VBZ	B-VP	*only*	RB	B-NP
the	DT	B-NP	*£*	#	I-NP
current	JJ	I-NP	*1.8*	CD	I-NP
account	NN	I-NP	*billion*	CD	I-NP
deficit	NN	I-NP	*in*	IN	B-PP
will	MD	B-VP	*September*	NNP	B-NP
narrow	VB	I-VP	*.*	.	O

Noun groups (NP) are in red and verb groups (VP) are in blue.

Evaluation

There are different kinds of measures to evaluate the performance of machine learning techniques, for instance:

- Precision and recall in information retrieval and natural language processing;
- The *receiver operating characteristic* (ROC) in medicine.

	Positive examples: P	**Negative examples:** N
Classified as P	True positives: A	False positives: B
Classified as N	False negatives: C	True negatives: D

More on the receiver operating characteristic here: `http://en.wikipedia.org/wiki/Receiver_operating_characteristic`

Recall, Precision, and the F-Measure

The **accuracy** is $\frac{|A \cup D|}{|P \cup N|}$.

Recall measures how much relevant examples the system has classified correctly, for P:

$$\text{Recall} = \frac{|A|}{|A \cup C|}.$$

Precision is the accuracy of what has been returned, for P:

$$\text{Precision} = \frac{|A|}{|A \cup B|}.$$

Recall and precision are combined into the **F-measure**, which is defined as the harmonic mean of both numbers:

$$F = \frac{2 \cdot \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Evaluation: Accuracy, precision, and recall

For noun groups with the predicted output:

	Word	POS	Group	Predicted			Word	POS	Group	Predicted	
	He	PRP	B-NP	B-NP			*to*	TO	B-PP	B-PP	
	reckons	VBZ	B-VP	B-VP			*only*	RB	B-NP	B-NP	X
	the	DT	B-NP	B-NP	X		*£*	#	I-NP	I-NP	X
	current	JJ	I-NP	**B-NP**	X		*1.8*	CD	I-NP	**B-NP**	X
	account	NN	I-NP	I-NP	X		*billion*	CD	I-NP	I-NP	X
	deficit	NN	I-NP	I-NP	X		*in*	IN	B-PP	B-PP	
	will	MD	B-VP	B-VP			*September*	NNP	B-NP	B-NP	
	narrow	VB	I-VP	I-VP			*.*	.	O	O	

There are 16 chunk tags, 14 are correct: Accuracy = $\frac{14}{16} = 0.875$

There are 4 noun groups, the system retrieved 2 of them: Recall = $\frac{2}{4} = 0.5$

The system identified 6 noun groups, two are correct: Precision = $\frac{2}{6} = 0.33$

Harmonic mean = $2 \times \frac{0.33 \times 0.5}{0.33 + 0.5} = 0.4$

Message Understanding Conferences

The Message Understanding Conferences (MUCs) measure the performance of information extraction systems.
They are competitions organized by an agency of the US department of defense, the DARPA
The competitions have been held regularly until MUC-7 in 1997.
The performances improved dramatically in the beginning and stabilized then.
MUCs are divided into a set of tasks that have been changing over time.
The most basic task is to extract people and company names.
The most challenging one is referred to as information extraction.

Information Extraction

Information extraction consists of:

- The analysis of pieces of text ranging from one to two pages,
- The identification of entities or events of a specified type,
- The filling of a pre-defined template with relevant information from the text.

Information extraction then transforms free texts into tabulated information.

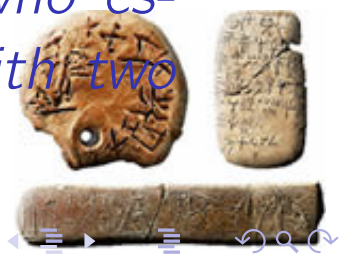
An Example

San Salvador, 19 Apr 89 (ACAN-EFE) – [TEXT] Salvadoran President-elect Alfredo Cristiani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime...

Garcia Alvarado, 56, was killed when a bomb placed by urban guerrillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador...

Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle...

According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.



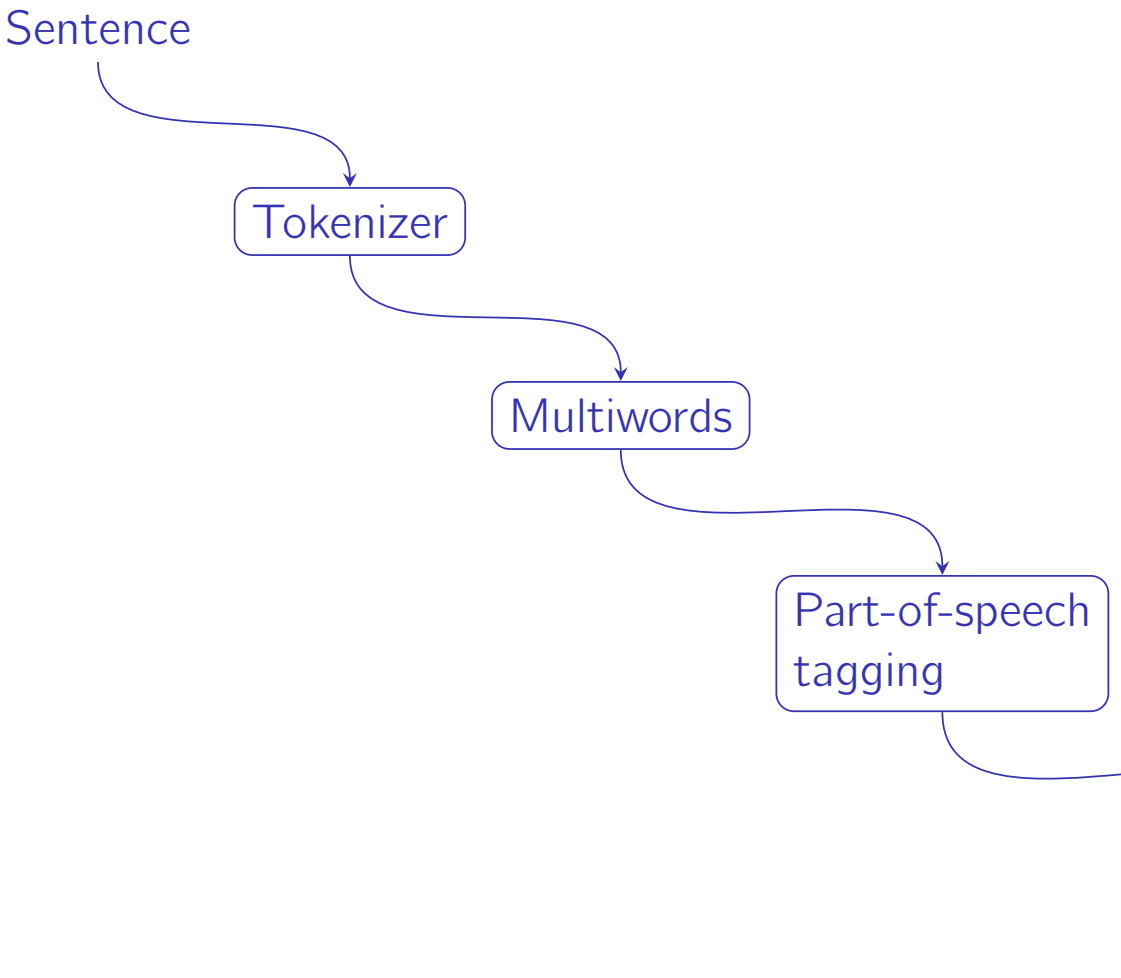
The Template

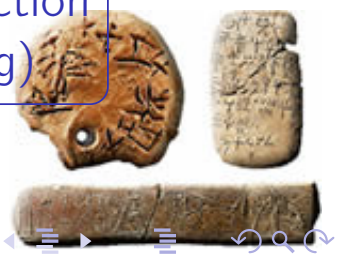
Template slots	Information extracted from the text
Incident: Date	19 Apr 89
Incident: Location	El Salvador: San Salvador (city)
Incident: Type	Bombing
Perpetrator: Individual ID	*urban guerrillas*
Perpetrator: Organization ID	*FMLN*
Perpetrator: Organization confidence	Suspected or accused by authorities: *FM*
Physical target: Description	*vehicle*
Physical target: Effect	Some damage: *vehicle*
Human target: Name	*Roberto Garcia Alvarado*
Human target: Description	*Attorney general*: *Roberto Garcia Alvarad* *driver* *bodyguards*
Human target: Effect	Death: *Roberto Garcia Alvarado* No injury: *driver* Injury: *bodyguards*

FASTUS

The FASTUS system has been designed at the Stanford Research Institute to extract information from free-running text
FASTUS uses partial parsers that are organized as a cascade of finite-state automata.
It includes a tokenizer, a multiword detector, and a group detector as first layers.
Verb groups are tagged with active, passive, gerund, and infinitive features.
Then FASTUS combines some groups into more complex phrases and uses extraction patterns to fill the template slots.

FASTUS' Architecture





Tokenization Revisited

- Some Asian languages do not include tokenization marks as in:
  然而，這樣的處理也衍生了一些問題。
  'However, this treatment also created some problems.'
  From Universaldependencies.org
- Tokenized as: 然而||，||這樣||的||處理||也||衍生||了||一些||問題||。
- Shao proposed the tokenization with the tagset: B, I, E, and S, where
  - B is the beginning of a word, I is inside, and E is the end.
  - S is for a single-character word.

然	而	，	這	樣	的	處	理	也	衍	生	了	一	些	問	題	。
B	E	S	B	E	S	B	E	S	B	E	S	B	E	B	E	S

Adaptation to Other Languages

In other languages, we have tokenization markers, mostly spaces.
We mark them with the X tag.
An example in French:

Chars: `On considère qu'environ 50 000 Allemands du Wartheland ont péri pendant la période.`
Tags: `BEXBIIIIIIIEXBIEBIIIIIEXBIIIIEXBIIIIIIIEXBEXBIIIIIIIIEXBIEXBIIEXBIIIIIEXBEXBIIIIIES`

Finally, we can use a final tag T to mark the end of a sentence.
This will enable us to carry out jointly tokenization and the sentence segmentation of a text.

Training the Model

The sentence # sent_id = test-s1
text = 然而，這樣的處理也衍生了一些問題。
The tokenized version from universal dependencies:

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	然而	然而	ADV	RB	_	7	mark	_	SpaceAfter=No
2	,	,	PUNCT	,	_	7	punct	_	SpaceAfter=No
3	這樣	這樣	PRON	PRD	_	5	det	_	SpaceAfter=No
4	的	的	PART	DEC	Case=Gen	3	case	_	SpaceAfter=No
5	處理	處理	NOUN	NN	_	7	nsubj	_	SpaceAfter=No
6	也	也	ADV	RB	_	7	mark	_	SpaceAfter=No
7	衍生	衍生	VERB	VV	_	0	root	_	SpaceAfter=No
8	了	了	AUX	AS	Aspect=Perf	7	aux	_	SpaceAfter=No
9	一些	一些	ADJ	JJ	_	10	amod	_	SpaceAfter=No
10	問題	問題	NOUN	NN	_	7	obj	_	SpaceAfter=No
11	。	。	PUNCT	.	_	7	punct	_	SpaceAfter=No