A brief note on features

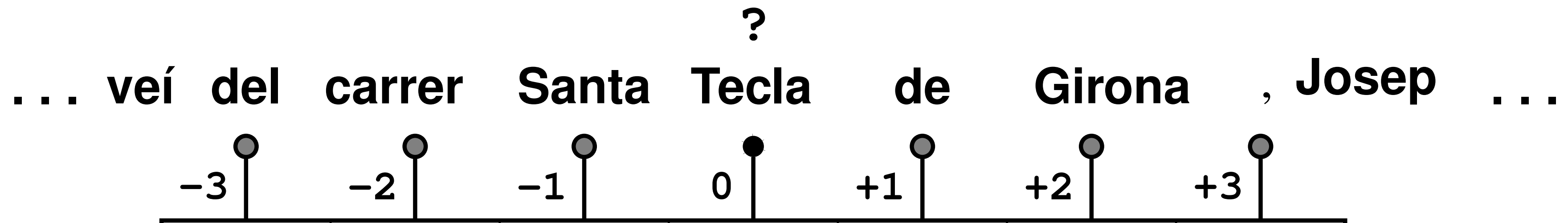
- The learning algorithm deals with a representation of training examples x
- Feature codification function: $\Phi : \mathcal{X} \longrightarrow \mathbb{R}^n$
- $\Phi(x)$ is a vector of features, with values in $\mathbb{R}$.
- Basic feature codification for NLP local decision problems follow the sliding window approach: codification of the local context.

Extracting features: Sliding Window

?

. . . veí del carrer Santa Tecla de Girona , Josep . . .

Extracting features: Sliding Window



Form	del	carrer	santa	tecla	de	girona	,
PoS	contr	n	adj	n	prep	n	,
Orto	min	min	Maj	Maj	min	Maj	punct
Prefix3	del	car	san	tec	de	gir	,
BIO	O	O	B				

Extracting features: Sliding Window

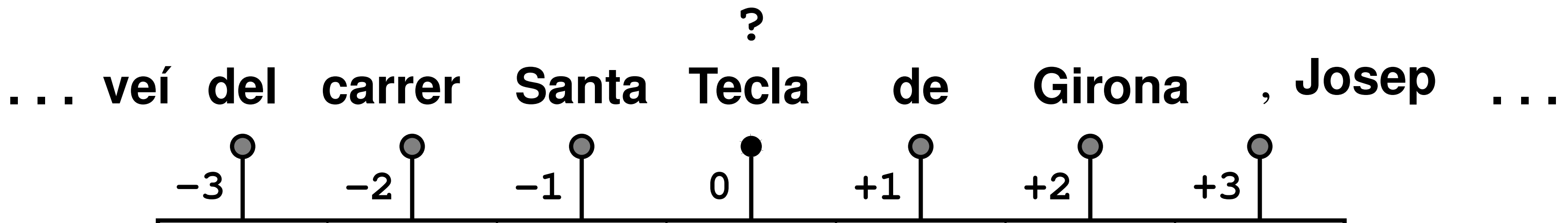
?

. . . veí del carrer Santa Tecla de Girona , Josep . . .

	−3	−2	−1	0	+1	+2	+3
Form	del	**carrer**	santa	tecla	de	girona	,
PoS	contr	n	adj	n	prep	n	,
Orto	min	min	Maj	Maj	min	Maj	punct
Prefix3	del	car	san	tec	de	gir	,
BIO	O	O	B				

`form@−2=carrer`

Extracting features: Sliding Window

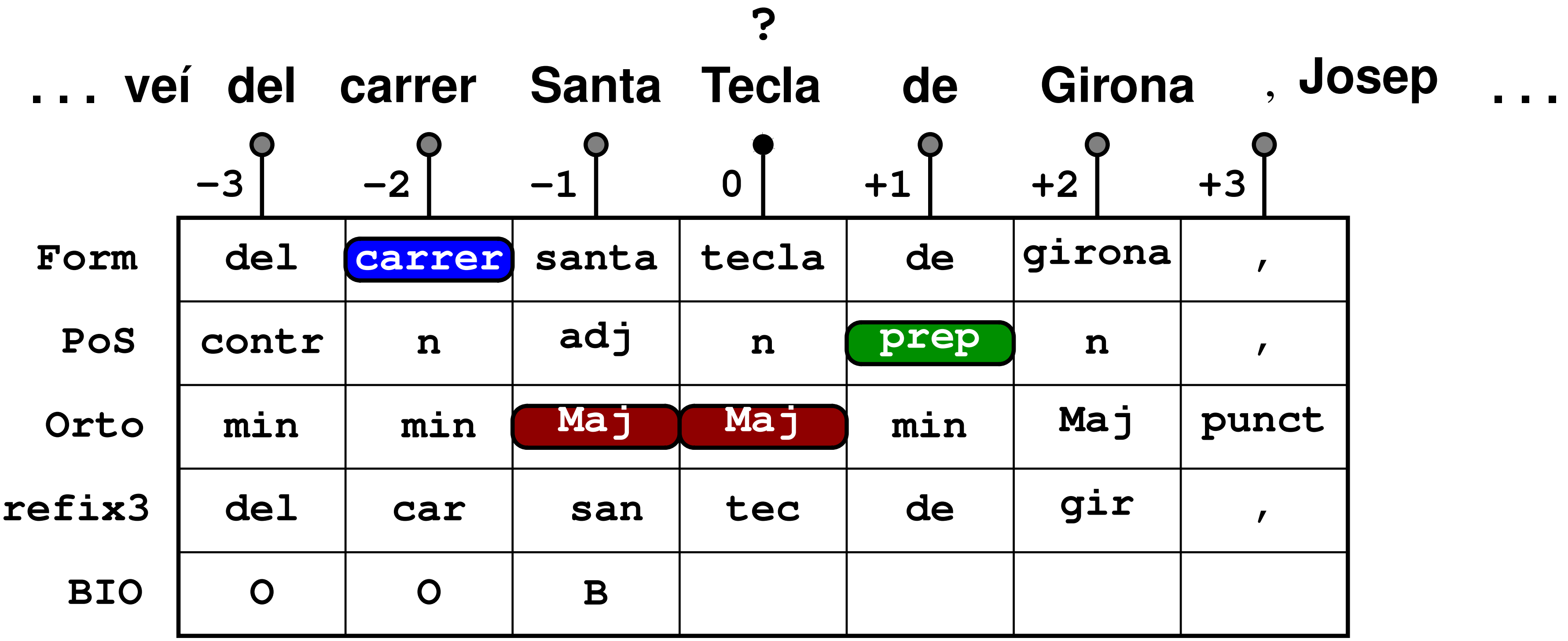


	−3	−2	−1	0	+1	+2	+3
Form	del	carrer	santa	tecla	de	girona	,
PoS	contr	n	adj	n	prep	n	,
Orto	min	min	Maj	Maj	min	Maj	punct
Prefix3	del	car	san	tec	de	gir	,
BIO	O	O	B				

`form@−2=carrer`

`pos@+1=prep`

Extracting features: Sliding Window



?

. . . veí del carrer Santa Tecla de Girona , Josep . . .

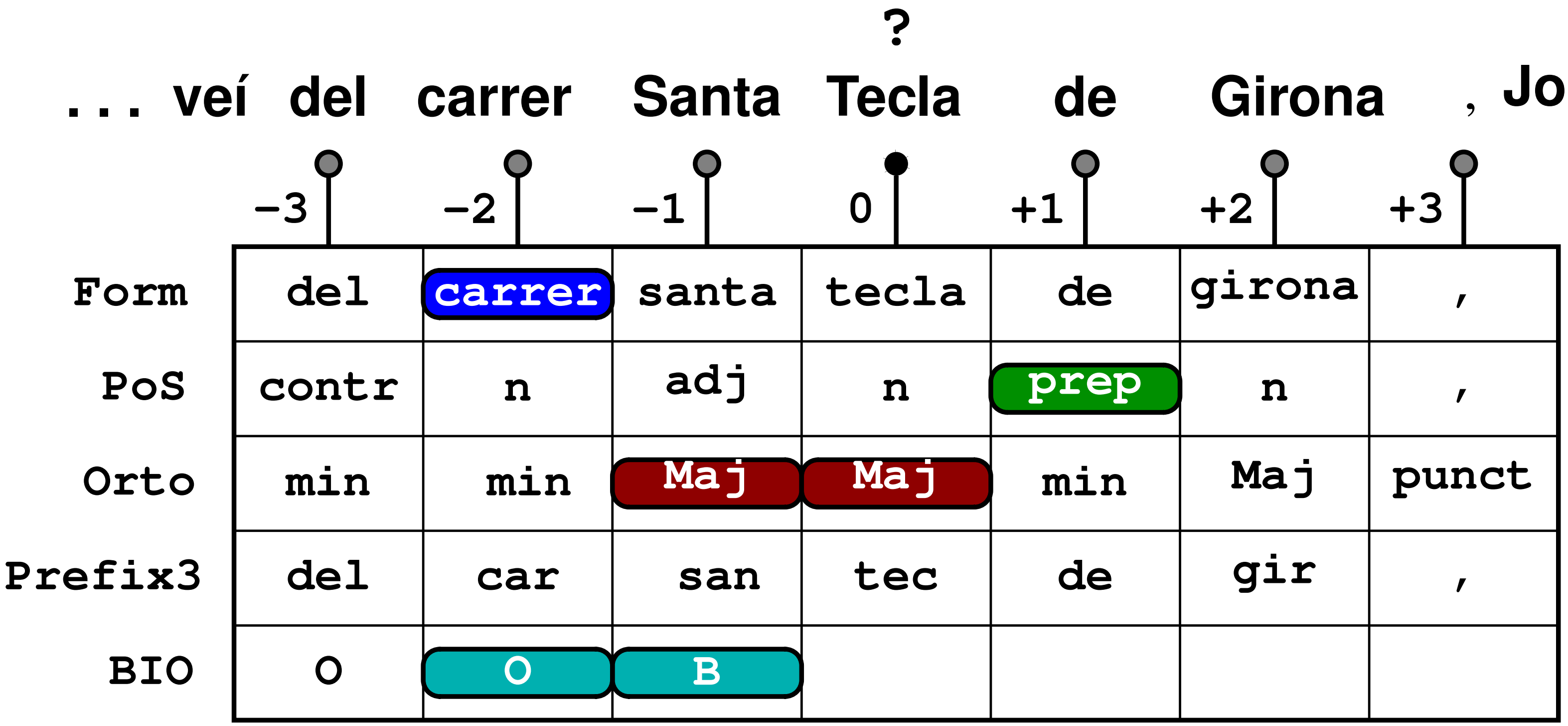
	−3	−2	−1	0	+1	+2	+3
Form	del	carrer	santa	tecla	de	girona	,
PoS	contr	n	adj	n	prep	n	,
Orto	min	min	Maj	Maj	min	Maj	punct
Prefix3	del	car	san	tec	de	gir	,
BIO	O	O	B				

`form@-2=carrer`

`pos@+1=prep`

`orto@-1:0=MajMaj`

Extracting features: Sliding Window



?

... veí del carrer Santa Tecla de Girona , Josep ...

	−3	−2	−1	0	+1	+2	+3
Form	del	carrer	santa	tecla	de	girona	,
PoS	contr	n	adj	n	prep	n	,
Orto	min	min	Maj	Maj	min	Maj	punct
Prefix3	del	car	san	tec	de	gir	,
BIO	O	O	B				

`form@−2=carrer` `orto@−1:0=MajMaj`

`pos@+1=prep` `bio@−2:−1=OB`

Snapshot of a codified example set$_{(1)}$

Confidence in the pound is widely expected to take another sharp dive ...

B-NP pos(+1):IN pos(-1):[OSB] pos(-1,0):[OSB]-NN pos(-1,0,+1):[OSB]-NN-IN pos(0):NN pos(0,+1):NN-IN w(+1):in w(-1):[OSB] w(-1,0):[OSB]-Confidence w(-1,0,+1):[OSB]-Confidence-in w(0):Confidence w(0,+1):Confidence-in
B-PP pos(+1):DT pos(-1):NN pos(-1,0):NN-IN pos(-1,0,+1):NN-IN-DT pos(0):IN pos(0,+1):IN-DT w(+1):the w(-1):Confidence w(-1,0):Confidence-in w(-1,0,+1):Confidence-in-the w(0):in w(0,+1):in-the
B-NP pos(+1):NN pos(-1):IN pos(-1,0):IN-DT pos(-1,0,+1):IN-DT-NN pos(0):DT pos(0,+1):DT-NN w(+1):pound w(-1):in w(-1,0):in-the w(-1,0,+1):in-the-pound w(0):the w(0,+1):the-pound
I-NP pos(+1):VBZ pos(-1):DT pos(-1,0):DT-NN pos(-1,0,+1):DT-NN-VBZ pos(0):NN pos(0,+1):NN-VBZ w(+1):is w(-1):the w(-1,0):the-pound w(-1,0,+1):the-pound-is w(0):pound w(0,+1):pound-is
B-VP pos(+1):RB pos(-1):NN pos(-1,0):NN-VBZ pos(-1,0,+1):NN-VBZ-RB pos(0):VBZ pos(0,+1):VBZ-RB w(+1):widely w(-1):pound w(-1,0):pound-is w(-1,0,+1):pound-is-widely w(0):is w(0,+1):is-widely
...

Snapshot of a codified example set$_{(2)}$

Numerical codification:

```
B-NP 1:1 6:1 31:1 33:1 41:1 84:1 559:1
B-PP 2:1 4:1 12:1 25:1 40:1 48:1 86:1 117:1 244:1
B-NP 3:1 5:1 11:1 27:1 29:1 47:1 83:1 85:1 243:1 2563:1 5741:1
I-NP 1:1 10:1 26:1 28:1 77:1 183:1 194:1 374:1 2714:1 5295:1
B-VP 2:1 68:1 76:1 185:1 192:1 420:1 774:1 2617:1 6501:1
I-VP 58:1 74:1 75:1 184:1 415:1 432:1 1214:1 1545:1 7769:1
I-VP 57:1 64:1 65:1 73:1 399:1 427:1 1251:1 2108:1 2827:1 6849:1
I-VP 56:1 63:1 67:1 72:1 100:1 396:1 547:1 1230:1 2022:1 2062:1 4471:1
I-VP 12:1 55:1 62:1 66:1 99:1 232:1 368:1 2102:1 2209:1 4234:1
B-NP 11:1 15:1 54:1 82:1 230:1 763:1 2056:1 2357:1 3362:1
...
```

Snapshot of a codified example set$_{(3)}$

Through a “dictionary” of features:

1 pos(0):NN 3208
2 pos(-1):NN 3206
3 pos(+1):NN 3168
4 pos(0):IN 2307
5 pos(-1):IN 2305
6 pos(+1):IN 2196
7 pos(0):NNP 2064
8 pos(-1):NNP 2062
9 pos(+1):NNP 1876
10 pos(-1):DT 1859
...
28 pos(-1,0):DT-NN 956
29 pos(0,+1):DT-NN 956
...