COMP4650 / COMP6490
Document Analysis 2018

Information Extraction

Gabriela Ferraro

Australian
National
University

Sequence labeling II

Weakness of Markov approaches in that it limits the context from which information can be extracted

Anything outside the context window has no impact on the decision being made...

* Acknowledgement: Some of the content originates from the Stanford NLP course at Coursera.org

Sequence labeling II

CRFs are indeed basically the sequential version of logistic regression

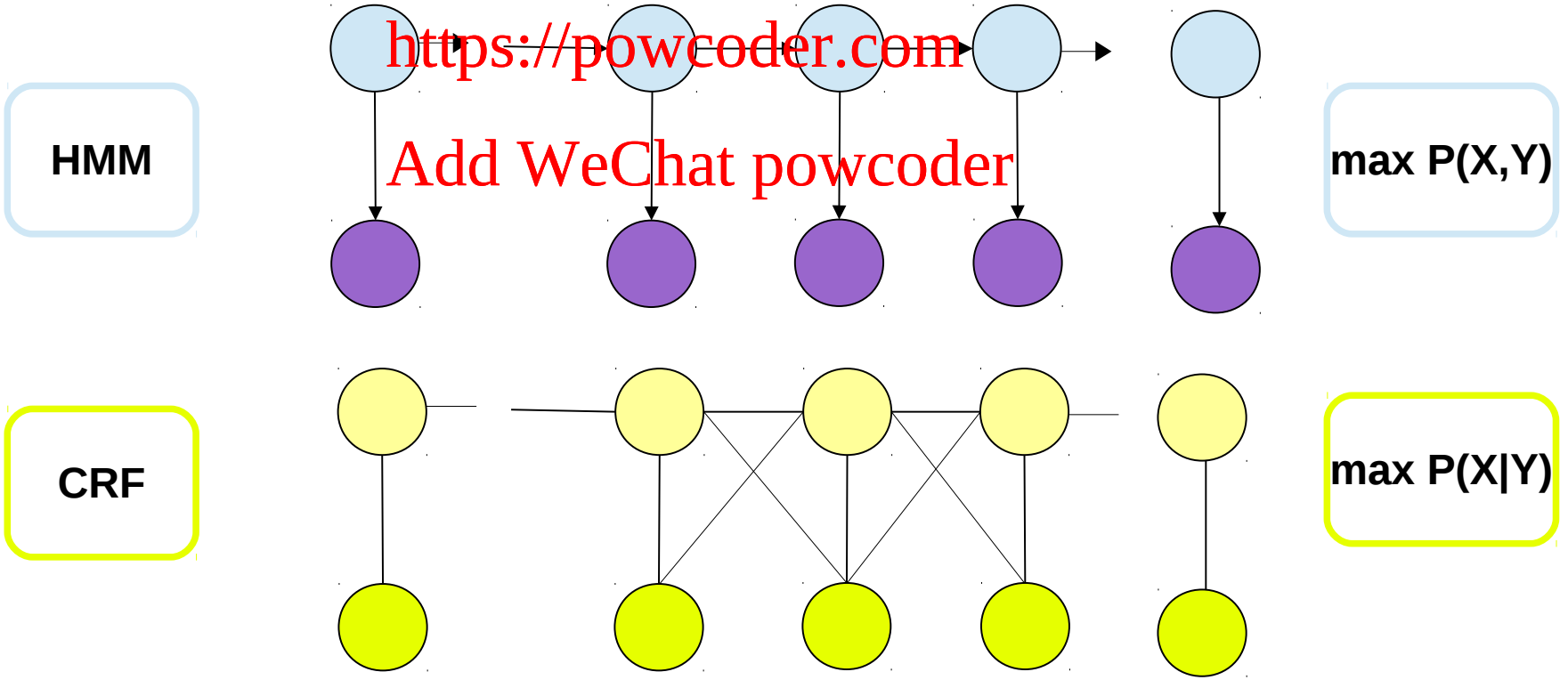
… whereas logistic regression is a log-linear model for classification, **CRFs are a log-linear model for sequential labels**.

Sequence labeling II

CRFs can define a much larger set of features

HMMs are necessarily local in nature because they're constrained to binary transition and emission feature functions

which force each word to depend only on the current label and each label to depend only on the previous label



Sequence labeling II

How to label a sentence using CRF?

The naive way is to calculate $p(labelsequence)$ for every possible labeling l, and then choose the label that maximizes this probability.

However, this is intractable…

A better way is to realize that (linear-chain) CRFs satisfy an optimal substructure property that allows us to use a dynamic programming algorithm to find the optimal label, e.g., the Viterbi algorithm for HMMs.

Information Extraction

BIO enconding

(B) beginning

(I) inside

(O) other

2n + 1 tags, where *n* is the number of entity types

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

BIO enconding

Without the B tag IO tagging is unable to distinguish between two entities of the same type that are right next to each other.

Since this situation doesn't arise very often (usually there is at least some punctuation or other delimiator)

+ IO tagging may be sufficient

+ advantage of using only n + 1 tags

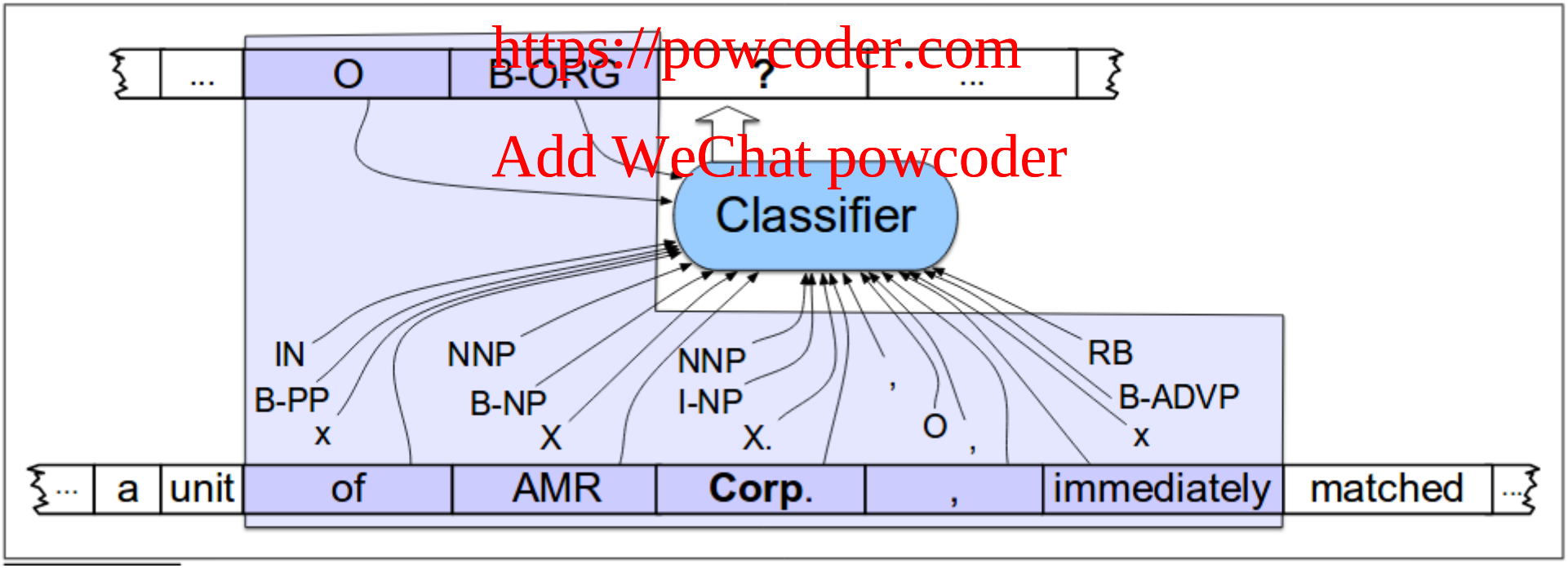
Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

Word-by-word feature encoding

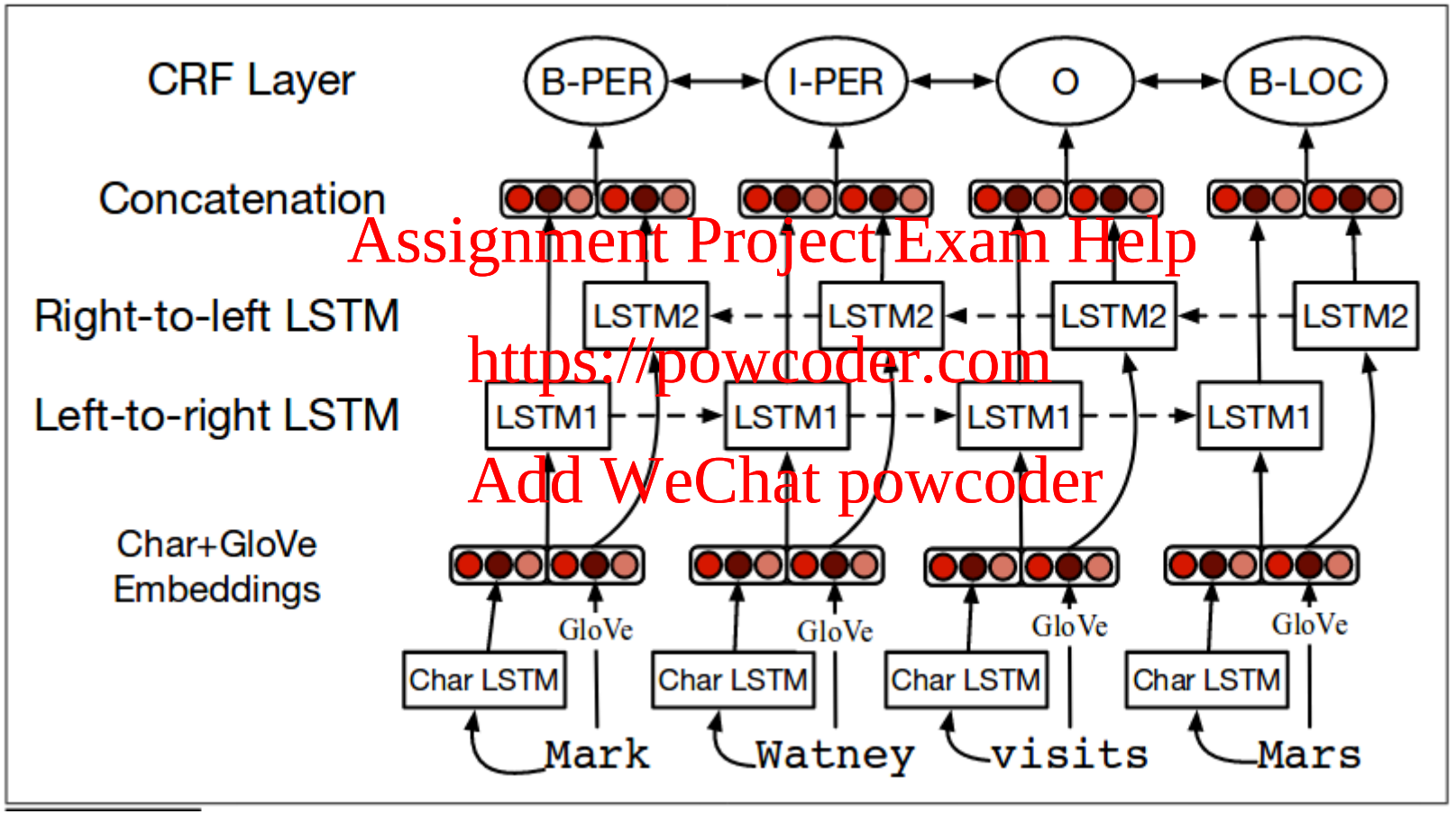
Word	POS	Chunk	Short shape	Label
American	NNP	B-NP	Xx	B-ORG
Airlines	NNPS	I-NP	Xx	I-ORG
,	,	O	,	O
a	DT	B-NP	x	O
unit	NN	I-NP	x	O
of	IN	B-PP	x	O
AMR	NNP	B-NP	X	B-ORG
Corp.	NNP	I-NP	Xx.	I-ORG
,	,	O	,	O
immediately	RB	B-ADVP	x	O
matched	VBD	B-VP	x	O
the	DT	B-NP	x	O
move	NN	I-NP	x	O
,	,	O	,	O
spokesman	NN	B-NP	x	O
Tim	NNP	I-NP	Xx	B-PER
Wagner	NNP	I-NP	Xx	I-PER
said	VBD	B-VP	x	O
.	,	O	.	O

Named Entity Recognition as sequence labeling

The features available to the classifier during training and classification are those in the boxed area



Neuronal algorithm for NER



- Use a CRF layer on top of the bi-LSTM
- Use Viterbi for decoding for selecting the most likely tag sequence

Information Extraction

- Extracting time and dates
  - Question answering
  - Calendar assistance
  - Personal assistance
  - …

Needs normalization!

So we can reason about them...

Extracting time and dates

- Absolute → map to calendar dates

- Relative → map to a particular time throught some other reference point

  - A week from last Tuesday.

- Duration → spans of time with differente granularities
  - seconds, minutes, days, weeks, centuries, etc

Extracting time and dates

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of '77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Extracting time and dates

Lexical triggers

Category	Examples
Noun	*morning, noon, night, winter, dusk, dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

A fare increase initiated <TIMEX3>last week</TIMEX3> by UALCorp's United Airlines was matched by competitors over <TIMEX3>the weekend</TIMEX3>, marking the second successful fare increase in<TIMEX3>two weeks</TIMEX3>.

(Pustejovskyet al. 2005, Ferro et al. 2005)

Extracting time and dates

Sequence labeling with BIO encoding

A fare increase initiated last week by UAL Corp's...

O O O O B I O O O

Feature	Explanation
Token	The target token to be labeled
Tokens in window	Bag of tokens in the window around a target
Shape	Character shape features
POS	Parts of speech of target and window words
Chunk tags	Base-phrase chunk tag for target and words in a window
Lexical triggers	Presence in a list of temporal terms

Extracting time and dates

Normalization: *VALUE attribute*

from the ISO 8601 standard for encoding temporal values

(ISO8601, 2004)

Unit	Pattern	Sample Value
Fully specified dates	YYYY-MM-DD	1991-09-28
Weeks	YYYY-Www	2007-W27
Weekends	PnWE	P1WE
24-hour clock times	HH:MM:SS	11:13:45
Dates and times	YYYY-MM-DDTHH:MM:SS	1991-09-28T11:00:00
Financial quarters	Qn	1999-Q3

Events Extraction

- Identify mentions of events in texts

  events can be assigned to point (or interval) in time

- sequence labeling

- BIO encoding

- usually applied supervised machine learning methods

Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character level suffixes for nominalizations (e.g., *-tion*)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
WordNet hypernyms	Hypernym set for the target

Events Extraction

Events + temporal expressions → Temporal ordering of the events

- Timeline

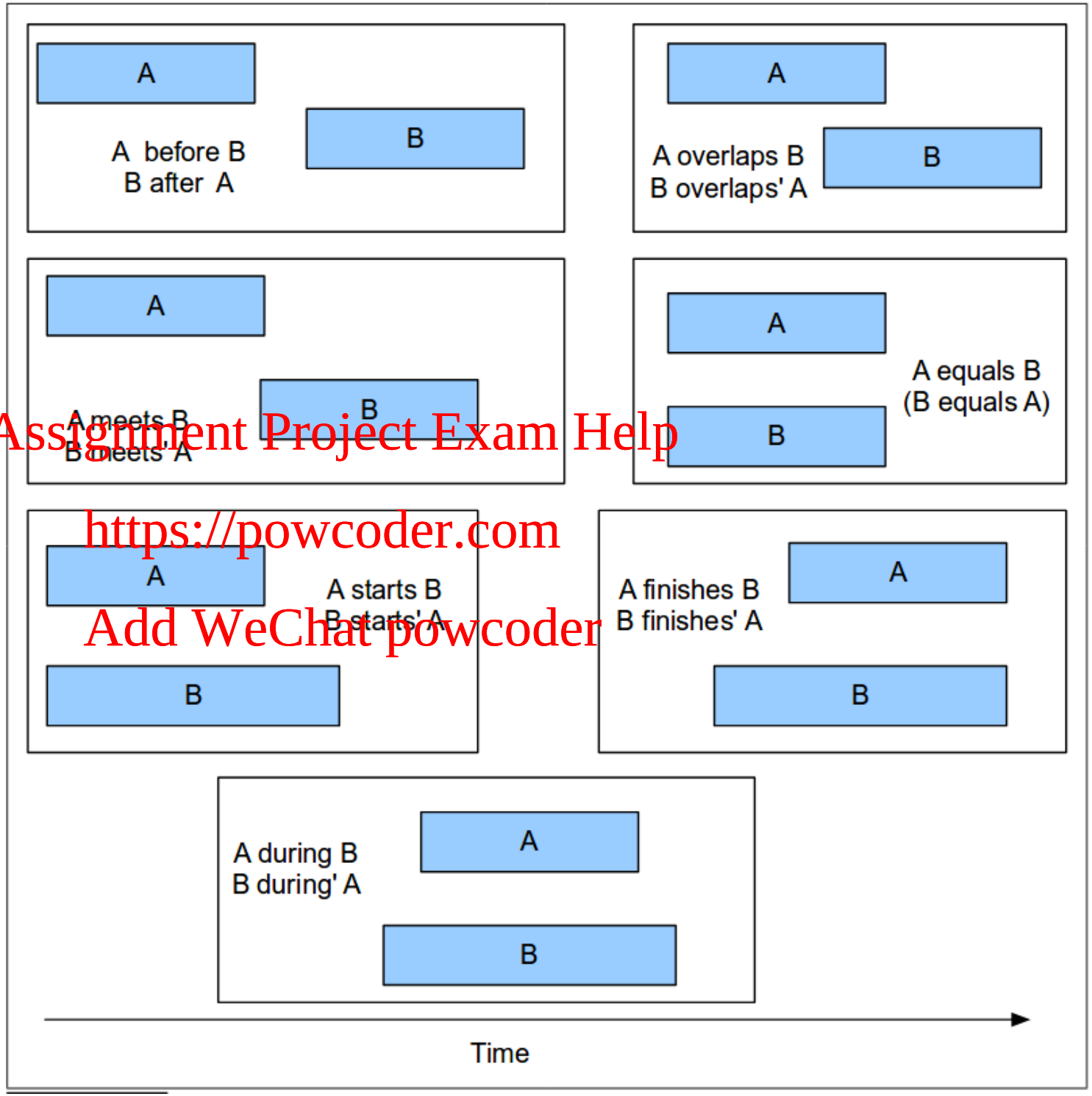
Classify events according to temporal relations

- Similar to Relation Extraction (instead of relations between entities, relations are between events)
- Finite set of temporal relations (Allen relations: Allen, 1984)

Useful for Q&A and summarization

TimeBank corpus

Allen relations between temporal events



Benchmarking

How do you know your method is working?

How good it is in respect to other methods?

Have a Baseline!!!

What is a baseline?

- Information that is used as a starting point by which to compare other information
- Benchmark
- Something you want to beat

How a baseline looks like?

- Random assignment
- Mayoririty class voting
- Simple heuristics
- Simple Machine Learning techniques
- Simple feature sets
- The system/method you want to beat!

Summary

- Named entities: `who and who's class (type)`
- Relation extraction: `who is doing what`
- Temporal expressions: `when? facilitate reasoning`
- Events: `facts`