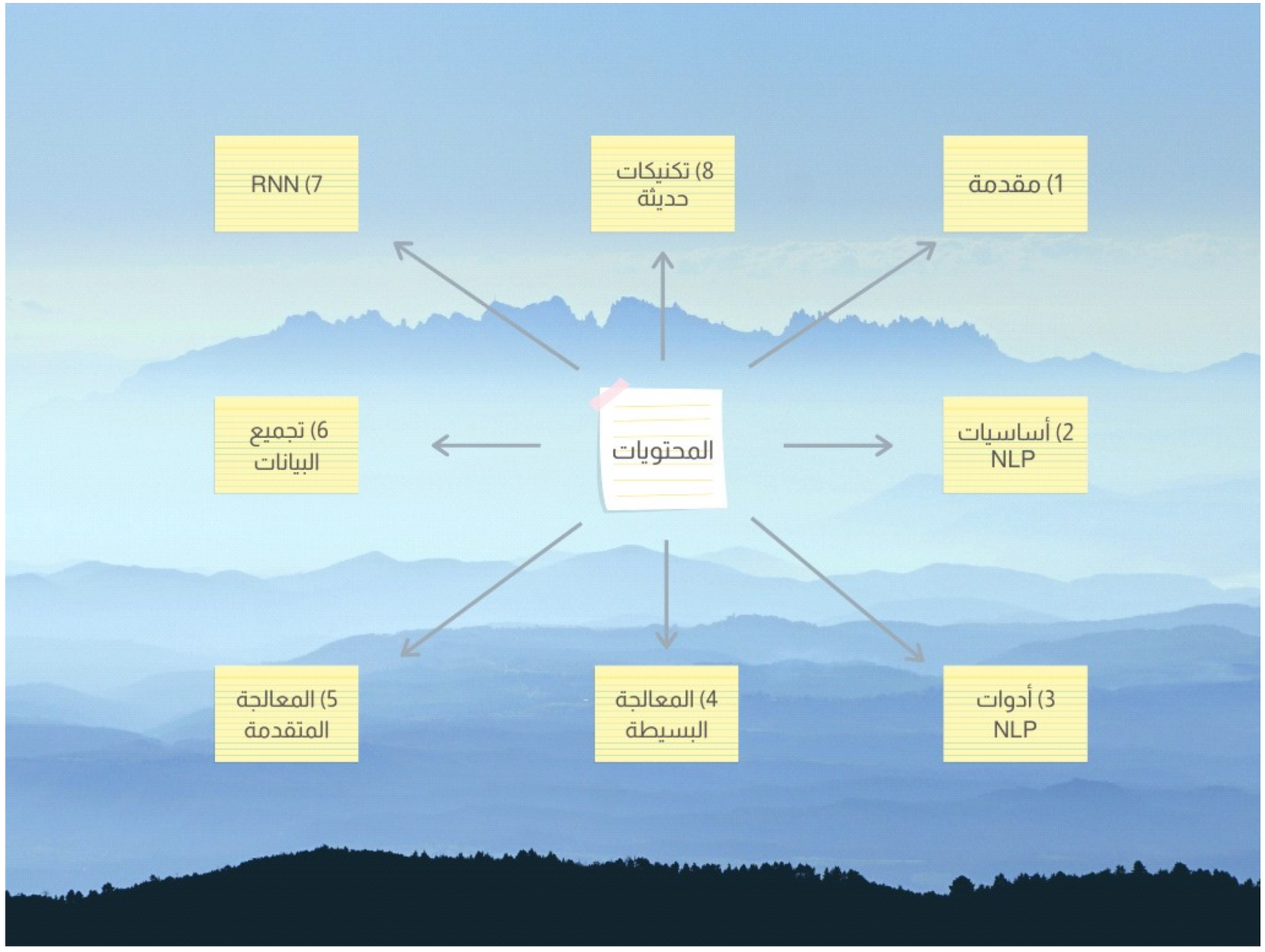
Natural Language Processing المعالجة اللغوية الطبيعية



المحتويات

1) مقدمة	المحتويات	ما هو NLP	تاريخ NLP	العقبات و التحديات	التطبيقات				
2) أساسيات NLP	المكتبات	الملفات النصية	ملفات pdf	البحث في النصوص					
3) أدوات NLP	Tokenization	Sent. Segm.	POS	Stem & Lemm	NER	Stopwords	Matchers	Syntactic Struc.	T.Visualization
4)المعالجة البسيطة	Word Meaning	Word embed	T. Vectors	Word2Vec	BOW	TF-IDF	Text Similarity	Dist. Similarity	
5)المعاجلة المتقدمة	T. Classification	T. Clustering	LDA	NMF	GloVe	Lexicons	NGrams	L. Modeling	T. Generation
	Sent. Analysis	Naïve Bayes	Vader	Auto Correct		Ans. Questions		Summarization & Snippets	
6)تجميع البيانات	Tweet Collecting	Data Scraping	Information Extraction			Information Retrieval		Relative Extraction	Search Engine
RNN (7	Seq to Seq	LSTM	GRU	Rec NN\TNN					
8)تكنيكات حديثة	Word Cloud	CNN	T. Forcing	Attention Model	Transformer	Bert	FastText	Gensim	Chat Bot

القسم السادس : تجميع البيانات

الجزء الثالث : Information Extraction

==========

نتناول الآن فرع هام من ال NLP و هو ما يسمي استخراج المعلومات Information Extraction

و هي القدرة علي استخراج المعلومات الهامة و البيانات المفيدة من النصوص , بشكل اوتوماتيك , بدون التدخل البشري . .

و هي لها فائدة كبيرة , في التعامل مع النصوص , حيث يتطلب في كثير من الاحيان استخراج المعلومات الهامة من النصوص , دون تضييع وقت في تحديدها

و أحيانا تكون هذه المعلومة المستخرجة هي علاقة بين متغيرين محددين فجملة **"المقر الرئيسي لشركة مرسيدس في برلين قد تعرض للاختراق"** , نعلم ان هناك ربط بين شركة مرسيدس , وانها المانية

او استخراج معلومة من الاساس **"الاقتصاد البريطاني قد يتعرض لهزة عنيفة عقب استفتاء البريكسيت "** فنعرف من النص ان الاقتصاد البريطاني , سيلحقه الأذي

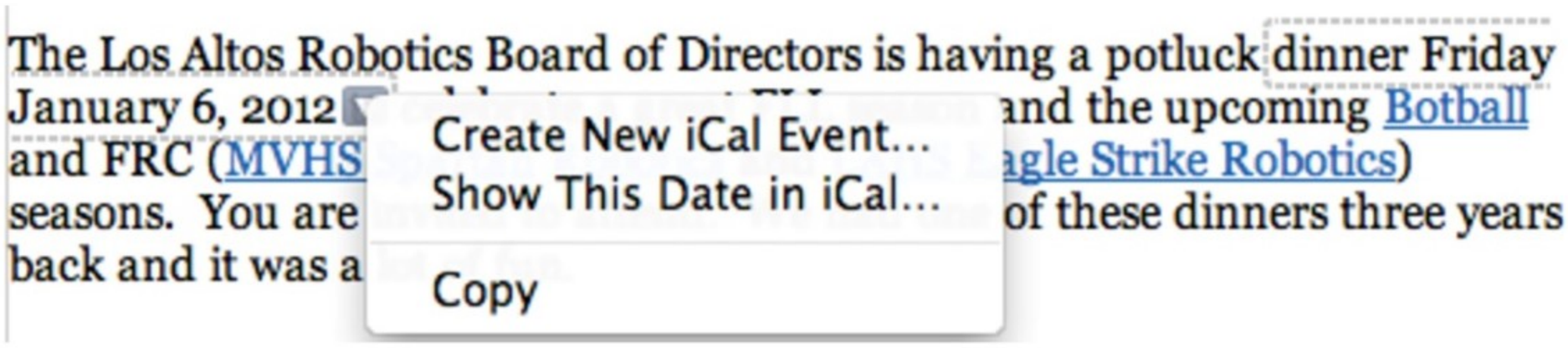
Information extraction (IE) systems

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information:
  - *relations* (in the database sense), a.k.a.,
  - a *knowledge base*
- Goals:
  1. Organize information so that it is useful to people
  2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

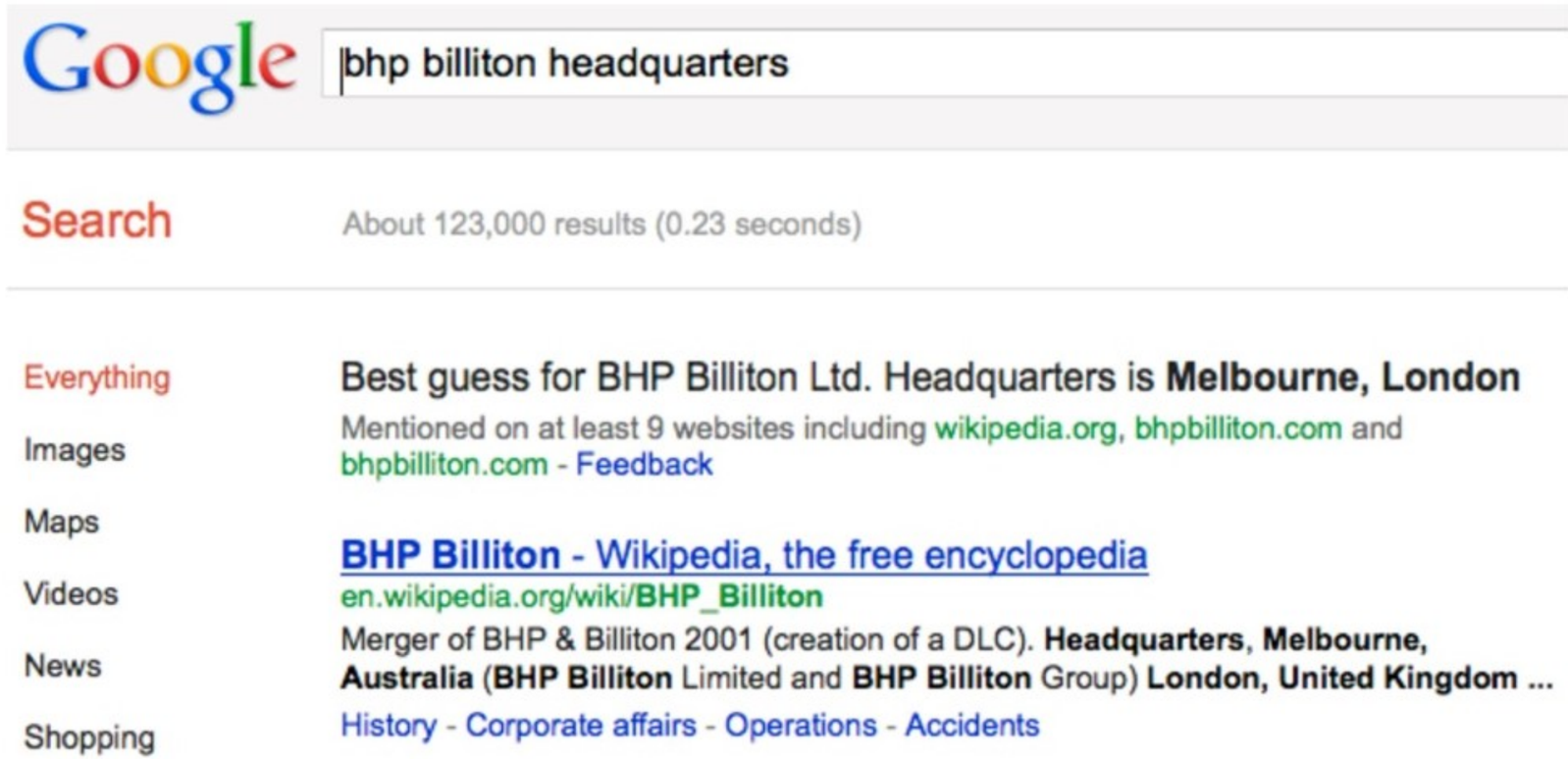
و كأن فكرة IE تقوم علي اجابة السؤال الهام : من فعل ماذا , ولمن و متي ؟
و ايضا من المطلوب استخراج معلومات عن اسماء الشركات و مقارها و مديريها و سياساتها و هكذا
ففي الجملة التالية , يمكن الربط بين كثير من المعلومات معا

- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- E.g.,
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
    - headquarters("BHP Biliton Limited", "Melbourne, Australia")
- Learn drug-gene product interactions from medical research literature

كما ان هناك تطبيقات ابسط و تتم بالفعل في العديد من خدمات الايميل , مثل ان يأتي موعد اجتماع , فيقوم ال calendar باقتراح أين تتم اضافته



كذلك لو بحثت عن اسم شركة , فستعلم جوجل ان هذه اسم شركة و هذه تفاصيلها



ومن اهم تطبيقاته ما يسمي NER اي تحديد اسماء الاعلام من اسماء الاشخاص و الشركات و المدن و العملات

_

نتناول هنا احد الخوارزميات المستخدمة في هذا الأمر هو ما يسمي MEMM

و تقوم فكرة MEMM او maximum entropy markov models علي فكرة بيانات التتابع sequence data

فالنصوص بطبيعتها هي sequence data لانها معتمدة علي تتالي الكلمات بعضها البعض , سواء تعاملنا معها ككلمات او حروف . .

و يمكن ان يتم معالجتها بأحد الطرق . .

- إما باستخراج الـ POS لها و التعامل معها
- او استخراج NER
- او في بعض اللغات (مثل الصينية) يتم تحديد بداية الكلمة , فنجد ان كل B تدل علي بداية كلمة معينة , و اذا كان يليها عدد من I فهي باقي الكلمة
- او اذا كان لدينا عدد من الاسئلة و الإجابات , ان نقوم بتحديد الأسئلة من الإجابات

Sequence problems

- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences …
- We can think of our task as one of labeling each item

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

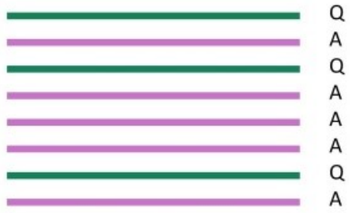
POS tagging

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Word segmentation



Text segmen-tation

و تعتمد فكرة MEMM علي أن الـ POS لا يكون قادر علي فهم جميع الكلمات الموجودة , فمثلا جملة
The Dow fell 22.6 %

ثم نتمكن من فهم اول 3 كلمات , ونعجز عن فهم 22.6 %

فيكون MEMM كأنه classifier لقراءة الفيتشرز و تحديد نوعها , و هنا نعتمد علي الفيتشرز المعتمدة علي الكلمة نفسها او الكلمات السابقة لها كما في الجدول الايمن

- For a Conditional Markov Model (CMM) a.k.a. a Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions
- A larger space of sequences is usually explored via search

Local Context

Decision Point

-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

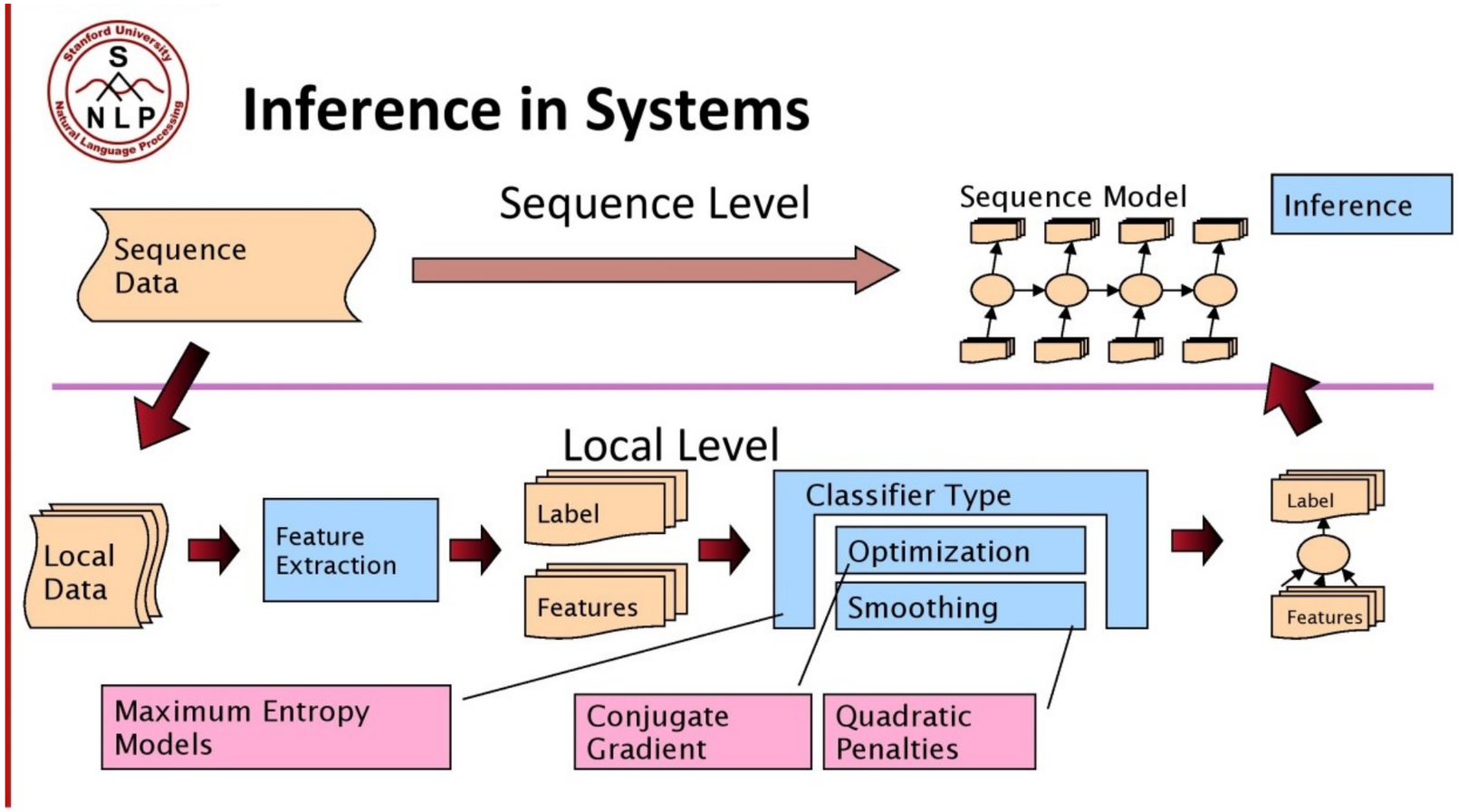
(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

Features

W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

فالفكرة قائمة علي التعامل مع sequence data و بالتالي بناء sequence model خاص بها

و بالتالي استخدام داتا لها , ثم عمل feature extraction و تحديد labels & features , ثم بناء ال classifier



_