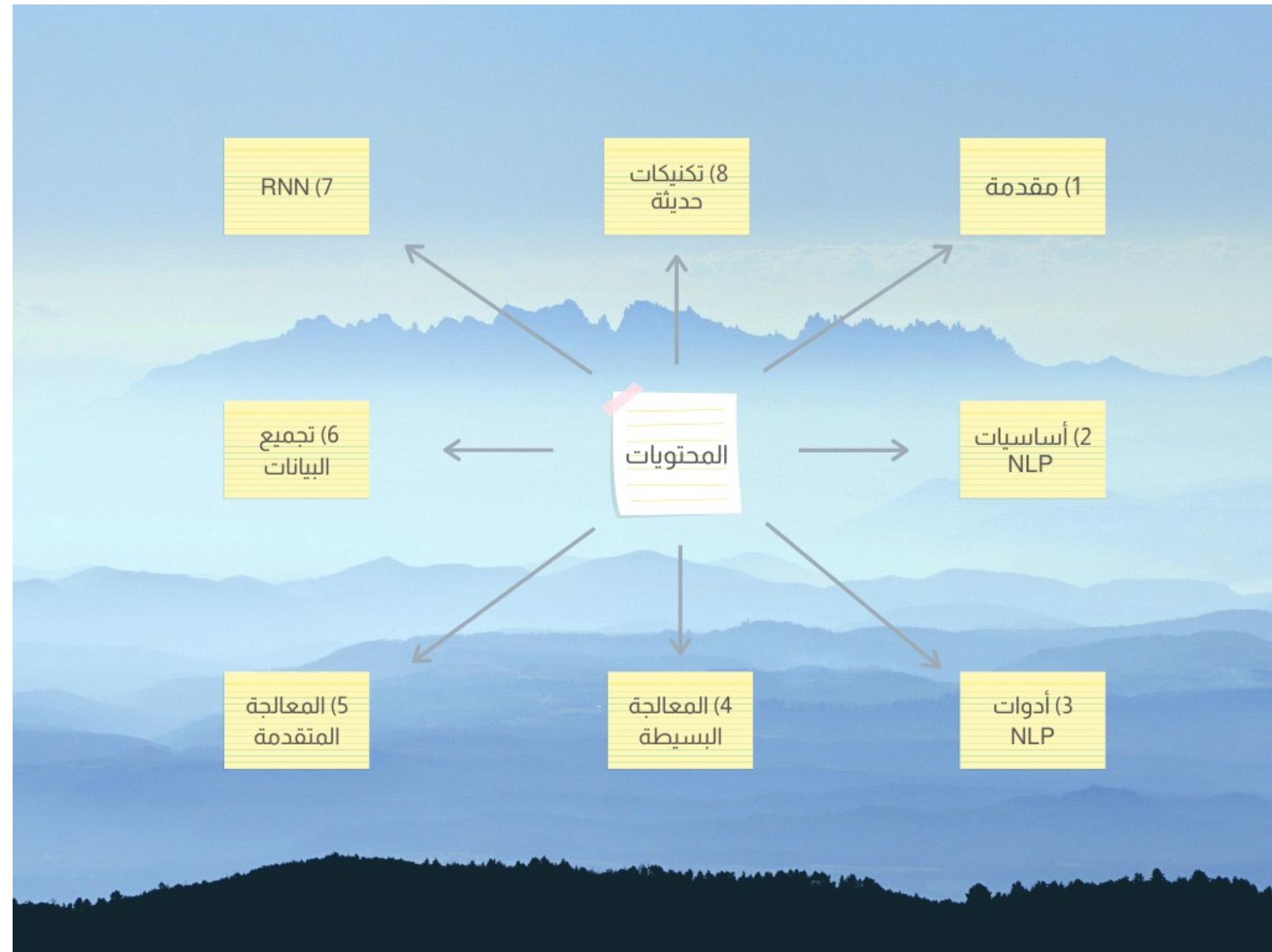


NATURAL LANGUAGE PROCESSING

المعالجة اللغوية الطبيعية



المحتويات

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

القسم السادس : تجميع البيانات

الجزء الخامس : Relative Extraction

نتحدث عن : استخراج العلاقات Relation Extraction

و هي طريقة استخراج العلاقات بين الكلمات بعضها البعض ، ومعرفة اي كلمة تتبع اي كلمة من حيث المعنى

فلو كان لدينا جملة معينة مثل هذه :

Company report: “International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)...”

فيتمكن استخراج العلاقات بين ان هناك شركة تسمى كذا ، و موقعها في مدينة كذا ، و تم انشائها عام كذا ، و اسمها الاصلي هو كذا

او قد يمكن استخراج معلومات ابسط من هذا

Extracted Complex Relation:

Company-Founding

Company	IBM
Location	New York
Date	June 16, 1911
Original-Name	Computing-Tabulating-Recording Co.

But we will focus on the simpler task of extracting relation **triples**

Founding-year(IBM,1911)

Founding-location(IBM,New York)

و يمكن فعل نفس الأمر من ويكيبيديا ، فعبر تصفح صفحة أحد الجامعات ، يمكننا معرفة الكثير عن علاقه الكلمة بكلمة و استخلاص الكثير من المعاني



The Leland Stanford Junior University, commonly referred to as Stanford University or Stanford, is an American private research university located in Stanford, California on the campus of 8,113 acres (3,285 ha) overlooking San Francisco Bay. It is a member of the Association of American Universities and is one of the most prominent, well-known, and influential research universities in the world. Stanford is a leading center of innovation and entrepreneurship, and its faculty and students have won numerous Nobel Prizes, Fields Medals, and Pulitzer Prizes. The university has been ranked among the top 10 in the world in various international rankings for over two decades. Stanford is also a major center for scientific research, particularly in the fields of engineering, computer science, and environmental science. The university is known for its commitment to diversity and inclusion, and it has received numerous awards for its work in these areas.

Stanford EQ Leland Stanford Junior University
Stanford LOC-IN California
Stanford IS-A research university
Stanford LOC-NEAR Palo Alto
Stanford FOUNDED-IN 1891
Stanford FOUNDER Leland Stanford

و تكون المعلومات مخزنة في ويكيبيديا على هيئة ملفات RDF ، وهنا أكثر من مليار ملف هكذا ، أكثر من ثلثها بالإنجليزية

Wikipedia Infobox		
Relations extracted from Infobox		
Stanford state California		
Type	Private	Stanford motto "Die Luft der Freiheit weht"
Endowment	US\$ 16.5 billion (2011) ^[3]	
President	John L. Hennessy	
Provost	John Etchemendy	
Academic staff	1,910 ^[4]	
Students	15,319	
Undergraduates	6,878 ^[5]	
Postgraduates	8,441 ^[5]	
Location	Stanford, California, U.S.	
Campus	Suburban, 8,180 acres (3,310 ha) ^[6]	
Colors	Cardinal red and white	

- Resource Description Framework (RDF) triples
 - subject predicate object
 - Golden Gate Park `location` San Francisco
 - `dbpedia:Golden_Gate_Park` `dbpedia-owl:location` `dbpedia:San_Francisco`
 - DBPedia: 1 billion RDF triples, 385 from English Wikipedia
 - Frequent Freebase relations:

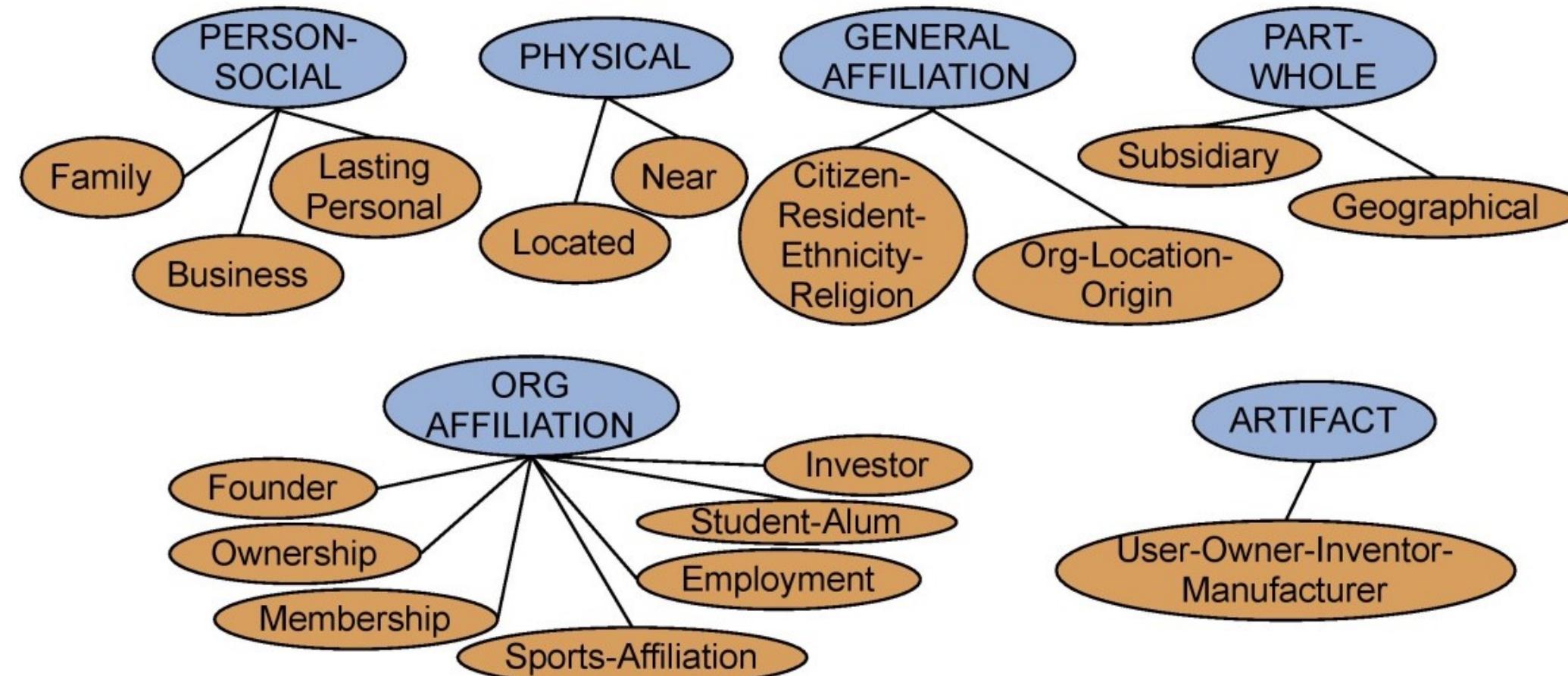
people/person/nationality,	location/location/contains
people/person/profession,	people/person/place-of-birth
biology/organism_higher_classification	film/film/genre

و هناك نموذج يسمى استخراج المحتوى التلقائي ACE , و الذي يقوم على 6 تقسيمات اساسية و 17 تقسيما فرعيا , بحيث يمكن من ايجاد العلاقة بين التقسيمات الفرعية و الاساسية بعضها البعض , بحيث عبر قراءة مقال ما , نتمكن من معرفة 3 معلومات مهمة عن الشخص , و 6 معلومات عن المؤسسة و هكذا , و بالطبع هذا معتمد على مدى توافر هذه البيانات في قلب الكلام



Automated Content Extraction (ACE)

17 relations from 2008 “Relation Extraction Task”



و هنا مثال , كيف ان جملة he was in Tennessee قامت بعمل ربط بين PER وهو person و GPE و هي تمكنا من ايجاد العلاقة بينهما , Geo Political Entity

- Physical-Located PER-GPE
He was in Tennessee
- Part-Whole-Subsidiary ORG-ORG
XYZ, the parent company of ABC
- Person-Social-Family PER-PER
John's wife Yoko
- Org-AFF-Founder PER-ORG
Steve Jobs, co-founder of Apple...

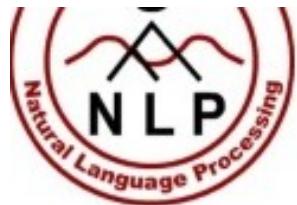
كما ان نموذج UMLS اعتمد على بيانات طبية لايجاد علاقات بين امراض و مسبباتها و اعراضها و طرق علاجها و هكذا



UMLS: Unified Medical Language System

- 134 entity types, 54 relations

Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function



Extracting UMLS relations from a sentence

Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes



Echocardiography, Doppler **DIAGNOSES** Acquired stenosis

كما ان هناك نموذج "علاقة الأصول" ، والذي يتمكن من معرفة اصل كلمة معينة ، فالزرافة كذا من الحيوانات و التي هي من الثدييات

و سان فرانسيسكو هي مدينة و تتبع ولاية كاليفورنيا التابعة للولايات المتحدة



Ontological relations

Examples from the WordNet Thesaurus

- IS-A (hypernym): subsumption between classes
 - Giraffe IS-A ruminant IS-A ungulate IS-A mammal IS-A vertebrate IS-A animal...
- Instance-of: relation between individual and class
 - San Francisco instance-of city

و يتم بناء الموديل اما عبر كود مخصص له

او Supervised ML او Semi Supervised او Unsupervised



How to build relation extractors

1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised and unsupervised
 - Bootstrapping (using seeds)
 - Distant supervision
 - Unsupervised learning from the web

--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

النماذج المستخدم لاستخراج العلاقات

و يمكن ان يتم بناء نماذج مخصصة ، لاستخراج العلاقات في المعلومات ، خاصة اذا كان هناك معلومات دائمة ما تأتي على هيئة نموذج محدد او قريب

فلو كان لدينا جملة طبية هكذا

- “Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”

يمكن ان نستخلص نموذج محدد وهو ان الكلمة التالية هي احد انواع الكلمة السابقة لها

red algae such as Gelidium

و بالتالي حينما تأتي الكلمة تكون الكلمة التالية احد انواع الكلمة السابقة such as

- “Y such as X ((, X)* (, and | or) X)”
- “such Y as X”
- “X or other Y”
- “X and other Y”
- “Y including X”
- “Y, especially X”

كذلك النماذج التالية لها ، تعتبر انواع من نماذج العلاقات بين الكلمات

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	... such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...

ايضا هناك كلمات مميزة تشير الي معنى محدد , مثل located in فالكلمة السابقة هي مؤسسة و التالية هي موقعها , وهكذا

- Intuition: relations often hold between specific entities
 - **located-in** (ORGANIZATION, LOCATION)
 - **founded** (PERSON, ORGANIZATION)
 - **cures** (DRUG, DISEASE)
- Start with Named Entity tags to help extract relation!

و هنا امثلة أخرى كيف يمكن استخراج علاقات بين المعلومات بعضها البعض

Who holds what office in what organization?

PERSON, POSITION of ORG

- George Marshall, Secretary of State of the United States

PERSON (named | appointed | chose | etc.) PERSON Prep? POSITION

- Truman appointed Marshall Secretary of State

PERSON [be]? (named | appointed | etc.) Prep? ORG POSITION

- George Marshall was named US Secretary of State

علي ان الامر ليس بهذه السهولة , فيمكن ان يكون هناك اسم مرض تالي لاسم دواء , دون ان نعلم حقيقة علاقته به , و كذلك امثلة أخرى



What relations hold between 2 entities?



PERSON

Founder?

Investor?

Member?

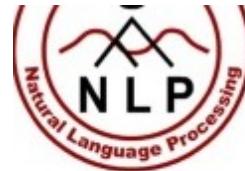
Employee?

President?



ORGANIZATION

و من مميزات هذا النوع انه بدقة عالية و يمكن تفصيله كما نشاء
لكن من عيوبه ان قيمة recall قليلة أي انه يضيع منه الكثير من المعلومات الحقيقية التي لن يراها , كما انها تحتاج الى
الكثير من العمل



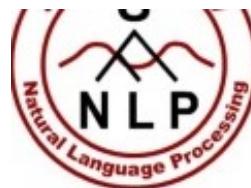
Hand-built patterns for relations

- Plus:
 - Human patterns tend to be high-precision
 - Can be tailored to specific domains
- Minus
 - Human patterns are often low-recall
 - A lot of work to think of all possible patterns!
 - Don't want to have to do this for every relation!
 - We'd like better accuracy

--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*

نناول الان فكرة عمل خوارزم بإشراف ليقوم بتحديد العلاقة بين الكلمات

- و تبدأ الخطوات تحديد البيانات التي سيتم استخراج المعلومات منها ، وتحديد الأسماء
- ثم عمل عنونة بشكل يدوي على الكلمات ذات العلاقات مع بعضها البعض
- ثم تدريب الخوارزم عليها



Supervised machine learning for relations

- Choose a set of relations we'd like to extract
- Choose a set of relevant named entities
- Find and label data
 - Choose a representative corpus
 - Label the named entities in the corpus
 - Hand-label the relations between these entities
 - Break into training, development, and test
- Train a classifier on the training set

و يقوم الخوارزم بعمل التالي :

- البحث لمعرفة هل هذين الكلمتين لها علاقة مع بعضهما البعض ام لا
- اذا وجد ان هناك علاقة , يبحث ليعرف مدى العلاقة
- و بعض الخوارزميات الاسرع تقوم بعمل حذف للكلمات التي بالتأكيد لا علاقة لها مع كلمات اخرى



How to do classification in supervised relation extraction

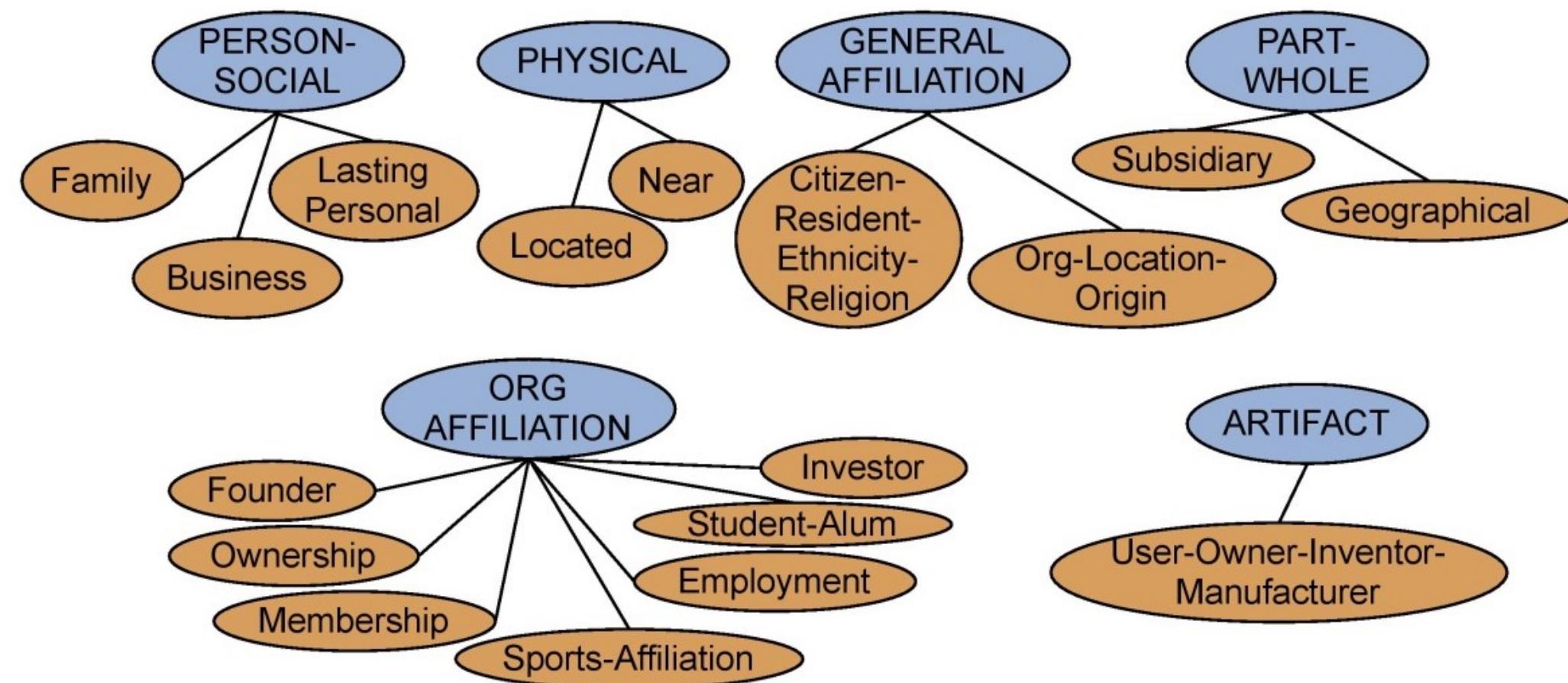
1. Find all pairs of named entities (usually in same sentence)
 2. Decide if 2 entities are related
 3. If yes, classify the relation
- Why the extra step?
 - Faster classification training by eliminating most pairs
 - Can use distinct feature-sets appropriate for each task.

و يتم الاعتماد على خريطة علاقات مثل خريطة ACE



Automated Content Extraction (ACE)

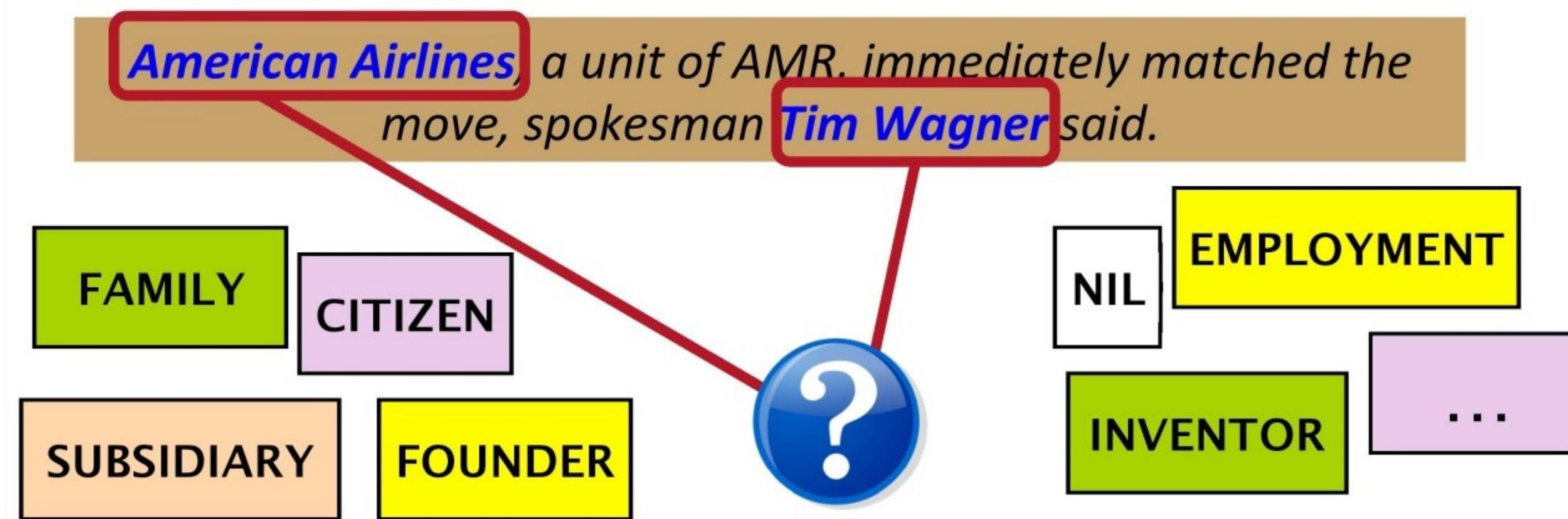
17 sub-relations of 6 relations from 2008 “Relation Extraction Task”



American Airlines a unit of AMR, immediately matched the move, spokesman Tim Wagner said

فيقوم الخوارزم اولاً بمعرفة ان هناك علاقة بين American Airlines و Tim Wagner ثم يبحث عن مدى العلاقة

Classify the relation between two entities in a sentence



و من الممكن استخراج عدد من الـ **features** المستخدمة لتحديد العلاقة مثل :

- الكلمات الاولى و الثانية و الجمع بينهم
 - عمل BOW لتشملهم جميعا
 - الكلمات السابقة و التالية للكلمة المطلوبة
 - الكلمات بينهم



Word Features for Relation Extraction

American Airlines, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said.

- Headwords of M1 and M2, and combination
Airlines Wagner Airlines-Wagner
 - Bag of words and bigrams in M1 and M2
{American, Airlines, Tim, Wagner, American Airlines, Tim Wagner}
 - Words or bigrams in particular positions left and right of M1/M2
M2: -1 *spokesman*
M2: +1 *said*
 - Bag of words or bigrams between the two entities
{a, AMR, of, immediately, matched, move, spokesman, the, unit}

- أيضا استخراج NER لهمها
- عمل دمج بين قيم NER
- تحديد القيمة الدقيقة لـ NER

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said

Mention 1	Mention 2
-----------	-----------

- Named-entity types
 - M1: ORG
 - M2: PERSON
- Concatenation of the two named-entity types
 - ORG-PERSON
- Entity Level of M1 and M2 (NAME, NOMINAL, PRONOUN)
 - M1: NAME [it or he would be PRONOUN]
 - M2: NAME [the company would be NOMINAL]

- تطبيق فكرة الـ parse , لمعرفة نوع الجملة الصغيرة التي يتكون منها الجمل الكبيرة (noun phrase , verb ..)
 - تحديد المسار الأعلى لكل جملة فيهم , اي الـ parent لها
 - تحديد علاقة الكلمات مع بعضها البعض



Parse Features for Relation Extraction

- Base syntactic chunk sequence from one to the other
NP NP PP VP NP NP
 - Constituent path through the tree from one to the other
NP ↑ NP ↑ S ↑ S ↓ NP
 - Dependency path
Airlines matched Wagner said

التعامل مع اصول الكلمات , فلو كان هناك ذكر لاسم مدينة , فيتم تحديد اسم المقاطعة و الولاية و الدولة , لو تم ذكر اسم مهنة يتم تحديد اسم المهنة الاكبر منها التي تشملها



Gazetteer and trigger word features for relation extraction

- Trigger list for family: kinship terms
 - parent, wife, husband, grandparent, etc. [from WordNet]
- Gazetteer:
 - Lists of useful geo or geopolitical words
 - Country name list
 - Other sub-entities

و يكون الشكل هكذا



American Airlines, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said.

Entity-based features

Entity ₁ type	ORG
Entity ₁ head	<i>airlines</i>
Entity ₂ type	PERS
Entity ₂ head	<i>Wagner</i>
Concatenated types	ORGPERS

Word-based features

Between-entity bag of words	{ <i>a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman</i> }
Word(s) before Entity ₁	NONE
Word(s) after Entity ₂	<i>said</i>

Syntactic features

Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	<i>Airlines</i> \leftarrow_{subj} <i>matched</i> \leftarrow_{comp} <i>said</i> \rightarrow_{subj} <i>Wagner</i>

ثم استخدام المصنف ، و حساب ادوات تقييم الموديل

و من مميزاته الدقة اذا توافرت الداتا الكافية

و من عيوبه التكلفة و الوقت ، و انه لن يتلائم الا مع داتا مشابهة ، فلو كنا نعرف اننا سنتعامل مع داتا مختلفة ، فلا يتم استخدامه

--*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*-*