

Natural Language Processing Applications



BITS Pilani
Pilani Campus

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 1

Date – 26th November 2023

These slides are prepared by the instructor, with grateful acknowledgement of James Allen and many others who made their course materials freely available online.

Session Content

- Objective of course
- What will we learn in this course?
- Text books and Reference books
- Evaluation Plan
- Application areas of Natural Language Processing



Objective of course

No	Course Objective
CO1	To provide students with the knowledge on designing and applying algorithms for real life NLP Applications
CO2	To give an overview of the major technologies used in NLP and hands-on experience of using such tools
CO3	To apply NLP techniques in state of art applications like Machine Translation, Information Extraction including Named entity recognition and Relation extraction
CO4	To give students opportunities to sharpen their programming skills for Natural Language Processing applications

What you will learn in this course

- **Grammar and spellcheckers**
- **Question Answering**
- **Chabot's**
- **Knowledge Graph Applications**
 - How to Use Knowledge Graphs
 - Why we need Knowledge Graphs
- **Machine Translation**
 - Statistical
 - Neural
 - Indic Languages
- **Information Extraction**
 - Named Entity Recognition
 - Relation Extraction
 - Extracting Events and Time
- **Sentiment Analysis**
 - Sentiment Analysis Methods
 - Rule based, ML based and Hybrid Systems
 - Neural Networks for Sentiment Analysis

Text books and Reference books

T1 Speech and Language processing: An introduction to Natural Language Processing, Computational Linguistics and speech Recognition by Daniel Jurafsky and James H. Martin

R1 Manning and Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA

R2 Neural Machine Translation by [Philip Koehn](#)

R3 Knowledge Graphs Methodology, Tools and Selected Use Cases by Dieter Fensel , Umutcan Şimşek, Kevin Angele, Elwin Huaman , Elias Kärle , Oleksandra Panasiuk , Ioan Toma, Jürgen Umbrich, and Alexander Wahler, Springer 2019

R4 Natural Language Toolkit. Bird and Loper, and other developers. Available for free at: – <http://www.nltk.org/>

Evaluation Plan

Name	Weight
Quiz (best 2 out of 3)	10%
Assignment 1 and 2	30%
Mid-term Exam	30%
End Semester Exam	30%

What is Natural Language Processing?

- Natural Language Processing
 - Process information contained in natural language text.
 - Also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)

NLP Tasks and Applications

It's a big world out there

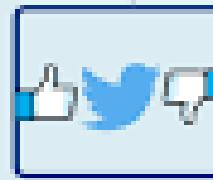


And everyone uses language

Information
Retrieval



Sentiment
Analysis



Information
Extraction



Machine
Translation

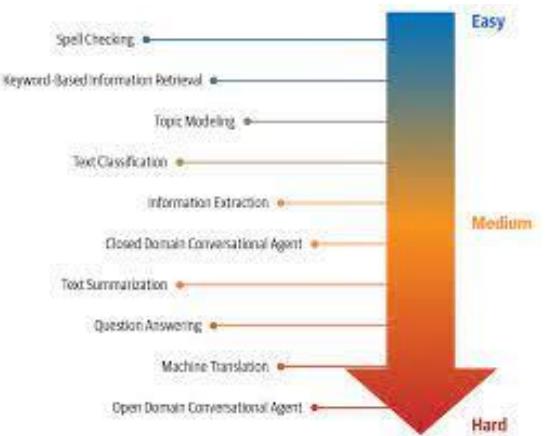
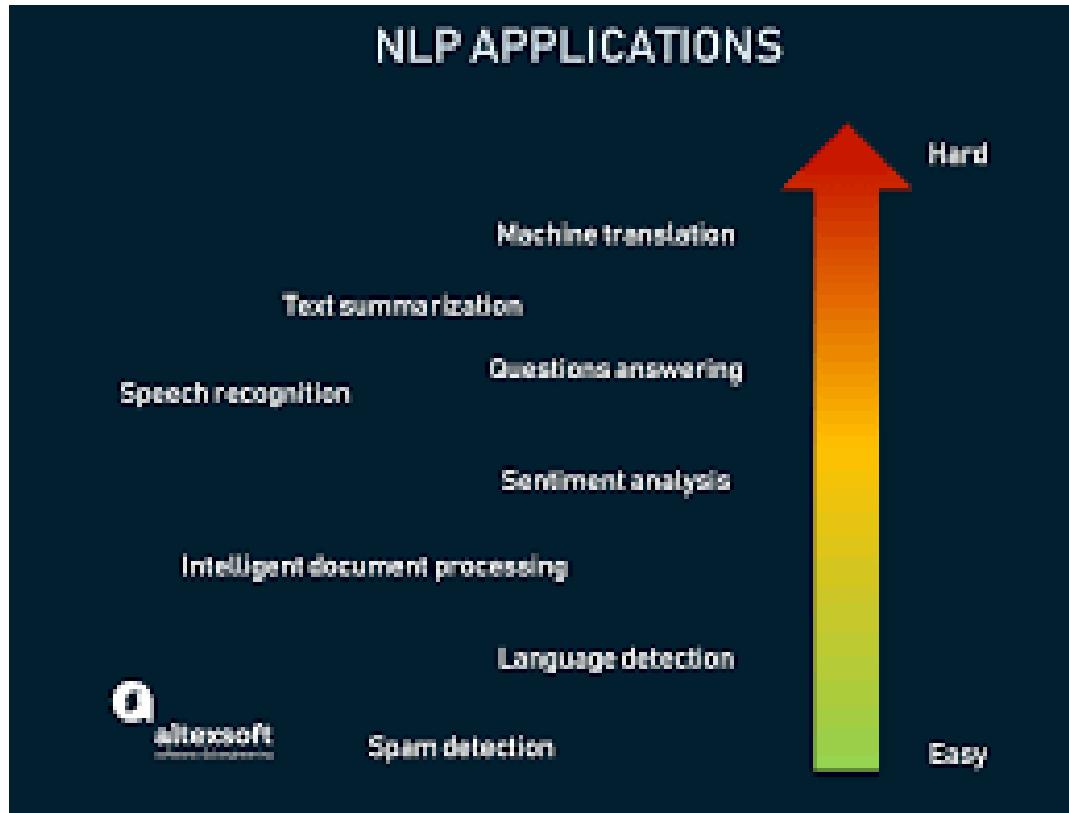


Natural Language Processing (NLP)

Question
Answering



NLP Applications and Difficulty Level



NLP Applications

- Question answering
 - Who is the first Taiwanese president?
 - Text Categorization/Routing
 - e.g., customer e-mails.
 - Text Mining
 - Find everything that can be done with NLP
 - Machine (Assisted) Translation
 - Language Teaching/Learning
 - Usage checking
 - Spelling correction
 - Is that just dictionary lookup?
-

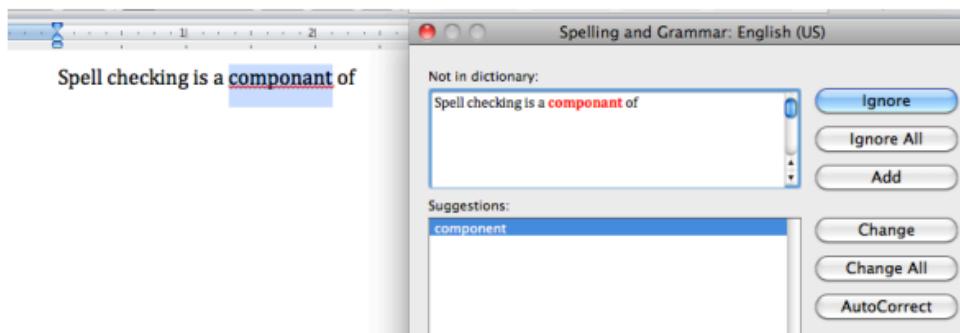
Application areas

- Text-to-Speech & Speech recognition
 - Healthcare
 - Natural Language Dialogue Interfaces to Databases
 - Information Retrieval
 - Information Extraction (<http://nlp.stanford.edu:8080/ner/process>)
 - Document Classification
 - Document Image Analysis
 - Automatic Summarization (<https://quillbot.com/summarize>)
 - Text Proof-reading – Spelling & Grammar
 - Machine Translation
 - **Fake News and Cyberbullying Detection**
 - **Monitoring Social Media Using NLP**
 - Plagiarism detection
 - Look-ahead typing / Word prediction_
 - Question Answering System (<http://start.csail.mit.edu/index.php>)
 - Sentiment Analysis (<https://komprehend.io/sentiment-analysis>)
-

Spell check

Applications for spelling correction

Word processing



Phones



Web search



2

Showing results for **natural language** processing
Search instead for natural langage processing

AI Assistants

Notification Assistant



Hi there - just a friendly reminder that your insurance policy expires in a month. Make sure to renew it in our member portal.

FAQ Assistant



I need to renew my renters insurance. How much will it be?



You can calculate your renewal price on our website here:
xyz.com/renew

Contextual Assistant



I need to renew my renters insurance. How much will it be?



I'd be happy to check for you. Firstly, are you still living in the same apartment?

Yes



Great - so just confirming it's 980 sq ft?

Yes



Thanks! Your new rate from September 1st onwards would be \$10 / month.



Would you like me to renew your policy for you right now?

Sure



Great. I've sent you a confirmation to your email.

AI Assistants

Personalized Assistant

- Assistant knows you much more in detail
- Quickly checks a few final things before giving you a quote tailored to your actual situation.



I can see your details are almost the same, except now you might want coverage for your new laptop. Additional coverage is only \$4 a month more for full coverage. Sound ok?

Sounds good!



Autonomous Organization of Assistants

- Group of AI assistants that know every customer personally
- Eventually run large parts of company operations—from lead generation over marketing, sales, HR, or finance



Knowledge Graph

• Knowledge Graph Applications

Applications

- Semantic Search
- Question Answering
- Analytics
- Dashboards
- Knowledge Sharing
- Knowledge Management

Algorithms

- Inferencing
- Machine Learning
- Entity Recognition
- Disambiguation
- Text Understanding
- Recommendations

Knowledge Graph



- Entities
- Relationships
- Semantic Descriptions

Data Sources

Data Transformation, Integration
Natural Language Processing



Machine Translation (MAT) – English to Hindi <http://anglahindi.iitk.ac.in>

Illustration of online MAT system

Simple Sentences.

sarala vaakya .

Welcome to London.

landana men aapaka svaagata hai.

There are some cases which are still pending.

vahaan kuc'ha kesa hain jo abhii bhii nilamibata hain .

.

Supervised Learning Methods

- Easy to build a “yes” or “no” predictor from supervised training data
 - Plenty of software packages to do the learning & prediction
 - Lots of people in NLP never go beyond this 😊
- Similarly, easy to build a system that chooses from a small finite set
 - Basically the same deal
 - But runtime goes up linearly with the size of the set
- Harder to predict the best string or tree (set is exponentially large or infinite)
 - Turn it into a sequence of predictions (words, tags, structure-building actions)
 - Tagging / finite-state / parsing algorithms will find the best sequence for you
 - Might also find k-best, or k-random (sample), or sum over all possible outputs

Text Annotation Tasks

1. Classify the entire document (“text categorization”)

Sentiment classification



What features of the text could help predict # of stars?
(e.g., using a log-linear model) How to identify more?
Are the features hard to compute? (syntax? sarcasm?)

★★★★★ **An extremely versatile machine!**, November 22, 2006

By [**Dr. Nickolas E. Jorgensen "njorgens3"**](#)

This review is from: Cuisinart DGB-600BC Grind & Brew, Brushed Chrome (Kitchen)

This coffee-maker does so much! It makes weak, watery coffee! It grinds beans if you want it to! It inexplicably floods the entire counter with half-brewed coffee when you aren't looking! Perhaps it could be used to irrigate crops... It is time-consuming to clean, but in fairness I should also point out that the stainless-steel thermal carafe is a durable item that has withstood being hurled onto the floor in rage several times. And if all these features weren't enough, it's pretty expensive too. If faced with the choice between having a car door repeatedly slamming into my genitalia and buying this coffee-maker, I'd unhesitatingly choose the Cuisinart! The coffee would be lousy, but at least I could still have children...

Other text categorization tasks



- Is it **spam**? (see [features](#))
- What **grade**, as an answer to this essay question?
- Is it **interesting to this user**?
 - News filtering; helpdesk routing
- Is it **interesting to this NLP program**?
 - Skill classification for a digital assistant!
 - If it's **Spanish**, translate it from Spanish
 - If it's **subjective**, run the sentiment classifier
 - If it's an **appointment**, run information extraction
- Where should it be **filed**?
 - Which mail folder? (work, friends, junk, urgent ...)
 - Yahoo! / Open Directory / digital libraries

Measuring Performance

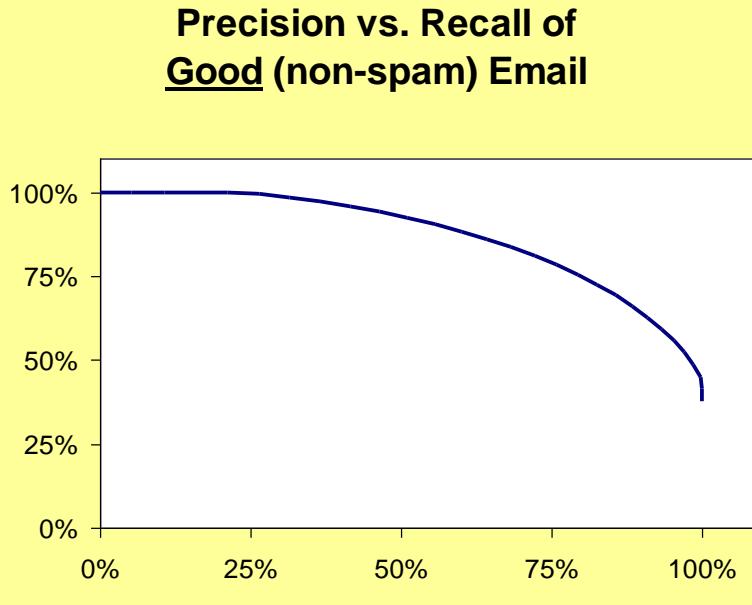


- Classification accuracy: What % of messages were classified correctly?
- Is this what we care about?

	Overall accuracy	Accuracy on spam	Accuracy on gen
System 1	95%	99.99%	90%
System 2	95%	90%	99.99%

- Which system do you prefer?

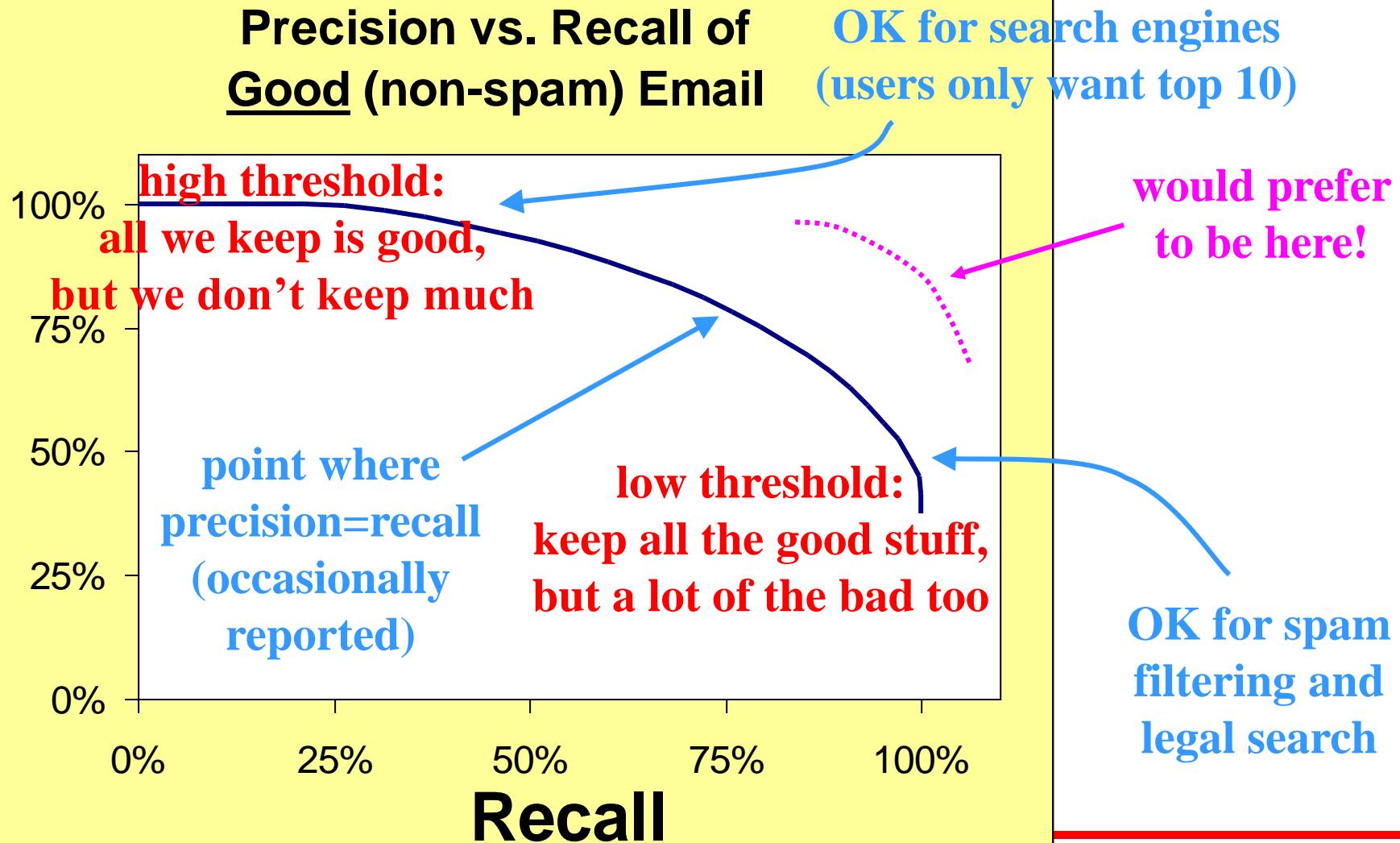
Measuring Performance



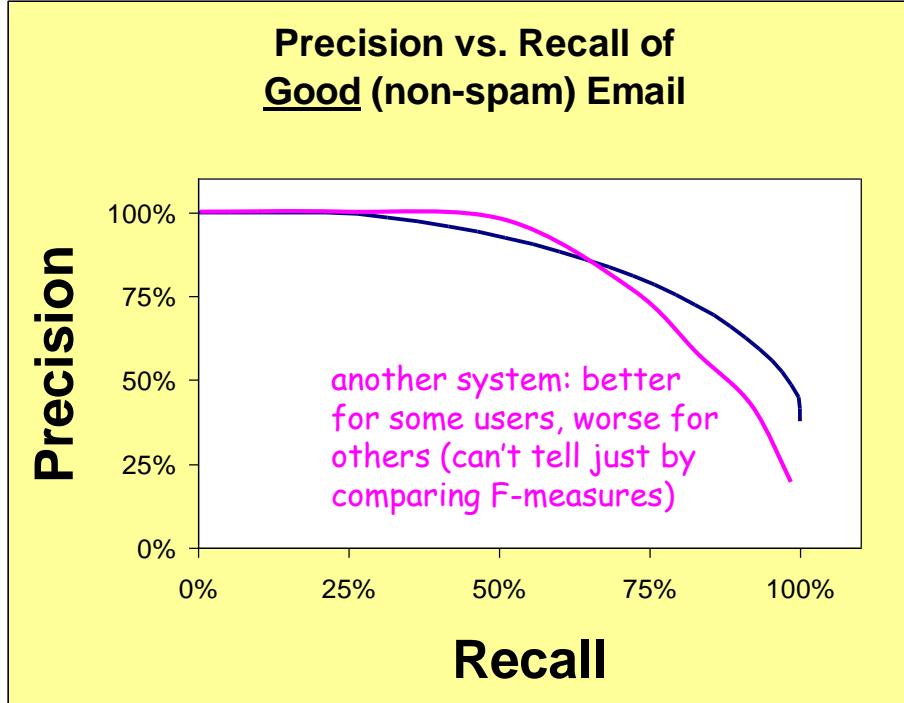
- **Precision** =
$$\frac{\text{good messages kept}}{\text{all messages kept}}$$
- **Recall** =
$$\frac{\text{good messages kept}}{\text{all good messages}}$$

Move from high precision to high recall by
deleting fewer messages (delete only if spamminess > high threshold)

Measuring Performance



Measuring Performance



- Precision =
$$\frac{\text{good messages kept}}{\text{all messages kept}}$$
- Recall =
$$\frac{\text{good messages kept}}{\text{all good messages}}$$
- F-measure =
$$\left(\frac{\text{precision}^{-1} + \text{recall}^{-1}}{2} \right)^{-1}$$

Move from high precision to high recall by deleting fewer messages (raise threshold)

Conventional to tune system and threshold to optimize F-measure on dev data
But it's more informative to report the whole curve

Since in real life, the user should be able to pick a tradeoff point they like

More than 2 classes

- Report F-measure for each class
- Show a confusion matrix

		Predicted class			
		Lion	Camel	Deer	
True class	Lion				
	Camel	I			
	Deer	I	I		

27

correct

Supervised Learning Methods

- **Nowadays:** Deep learning with pretraining
 - Usually transformers
- **Training:** Learn weights θ such that in training data, the *true* answer y^* tends to have a **high log-probability**
 - Or such that the decoder tends to have low loss
 - For a stochastic decoder, this means $\mathbb{E}_y[\text{loss}(y | y^*)]$ is low
- **Decoding:** Search for the highest-probability answer \hat{y}
 - Or an answer such that the Bayes risk $\mathbb{E}_y[\text{loss}(\hat{y} | y)]$ is low
- **Previously, used linear models with feature engineering**
 - Log-linear, perceptron, SVM, MIRA, ...
 - Feature engineering: Throw in enough features to fix most errors
 - Or to include conjunctions automatically: Kernel methods, decision forests
 - Learner just seeks weights θ such that in training data, the desired answer has a **higher score** than the wrong answers
 - If score → prob by softmax, then max log-prob of desired answer

Text Annotation Tasks

1. Classify the entire document
2. Classify individual word tokens

Word Sense Disambiguation (WSD)

Problem:

The company said the *plant* is still operating ...

- ⇒ (A) Manufacturing plant or
- ⇒ (B) Living plant

Training Data: Build a special classifier just for tokens of "plant"

Sense	Context
(1) Manufacturing " " " "	... union responses to <i>plant</i> closures
	... computer disk drive <i>plant</i> located in ...
	company manufacturing <i>plant</i> is in Orlando ...
(2) Living " " " "	... animal rather than <i>plant</i> tissues can be ...
	... to strain microscopic <i>plant</i> life from the ...
	and Golgi apparatus of <i>plant</i> and animal cells

Test Data:

Sense	Context
???	... vinyl chloride monomer <i>plant</i> , which is ...
???	... molecules found in <i>plant</i> tissue from the ...



WSD for Machine Translation (English → Spanish)

Problem:

... He wrote the last **sentence** two years later ...

⇒ *sentencia* (legal sentence) or

⇒ *frase* (grammatical sentence)

Training Data: Build a special classifier just for tokens of "sentence"

Translation	Context
(1) sentencia	... for a maximum <i>sentence</i> for a young offender ...
" "	... of the minimum <i>sentence</i> of seven years in jail ...
" "	... were under the <i>sentence</i> of death at that time ...
(2) frase	... read the second <i>sentence</i> because it is just as ...
" "	... The next <i>sentence</i> is a very important ...
" "	... It is the second <i>sentence</i> which I think is at ...

Test Data:

Translation	Context
???	... cannot criticize a <i>sentence</i> handed down by ...
???	... listen to this <i>sentence</i> uttered by a former ...



Accent Restoration in Spanish & French

Problem:

Input: ... déjà travaille cote a cote ...



Output: ... déjà travaillé côté à côté ...

Examples:

... appeler l'autre **cote** de l'atlantique ...

⇒ *côté* (meaning side) or

⇒ *côte* (meaning coast)

... une famille des **pecheurs** ...

⇒ *pêcheurs* (meaning fishermen) or

⇒ *pécheurs* (meaning sinners)



Accent Restoration in Spanish & French

Training Data:

Pattern	Context
(1) côté	... du laisser de <i>cote</i> faute de temps ...
” ”	... appeler l' autre <i>cote</i> de l' atlantique ...
” ”	... passe de notre <i>cote</i> de la frontiere ...
(2) côte	... vivre sur notre <i>cote</i> ouest toujours ...
” ”	... creer sur la <i>cote</i> du labrador des ...
” ”	travaillaient cote a <i>cote</i> , ils avaient ...

Test Data:

Pattern	Context
???	... passe de notre <i>cote</i> de la frontiere ...
???	... creer sur la <i>cote</i> du labrador des ...



Capitalization Restoration

Problem:

... FRIED CHICKEN, TURKEY SANDWICHES AND FROZEN ...

- ⇒ *turkey* (the *bird*) or
- ⇒ *Turkey* (the *country*)

Training Data:

Capitalization	Context
(1) turkey	... OF FRIED CHICKEN , TURKEY SANDWICHES AND FROZEN ...
” ”	... NTS A POUND , WHILE TURKEY PRICES ROSE 1.2 CENTS ...
” ”	... PLAY , REAL GRADE-A TURKEY , WHICH ONLY A PRICE ...
(2) Turkey	... INUNDATED EASTERN TURKEY AFTER THE EARLIER ...
” ”	... FEELINGS TOWARD TURKEY SURFACED WHEN GREECE ...
” ”	... THE CONTRACT WITH TURKEY WILL PROVIDE OPPORTU...

Test Data:

Capitalization	Context
???	... NECK LIKE THAT OF A TURKEY ON A CHOPPING BLOCK ...
???	... PROBLEM IS THAT TURKEY IS NOT A EUROPEAN ...



Text-to-Speech Synthesis

Problem:

... slightly elevated *lead* levels ...

⇒ *lɛd* (as in *lead mine*) or

⇒ *li:d* (as in *lead role*)

Training Data:

Pronunciation	Context
(1) lɛd " "	... it monitors the <i>lead</i> levels in drinking ...
	... conference on <i>lead</i> poisoning in ...
	... strontium and <i>lead</i> isotope zonation ...
(2) li:d " "	... maintained their <i>lead</i> Thursday over ...
	... to Boston and <i>lead</i> singer for Purple ...
	... Bush a 17-point <i>lead</i> in Texas , only 3 ...

Test Data:

Pronunciation	Context
???	... median blood <i>lead</i> concentration was ..
???	... his double-digit <i>lead</i> nationwide . The ...



Spelling Correction

Problem:

... and he fired presidential **aid/aide** Dick Morris after ...

⇒ *aid* or

⇒ *aide*

Training Data:

Spelling	Context
(1) aid " "	... and cut the foreign <i>aid/aide</i> budget in fiscal 1996 they offered federal <i>aid/aide</i> for flood-ravaged states ...
(2) aide " "	... fired presidential <i>aid/aide</i> Dick Morris after and said the chief <i>aid/aide</i> to Sen. Baker, Mr. John ...

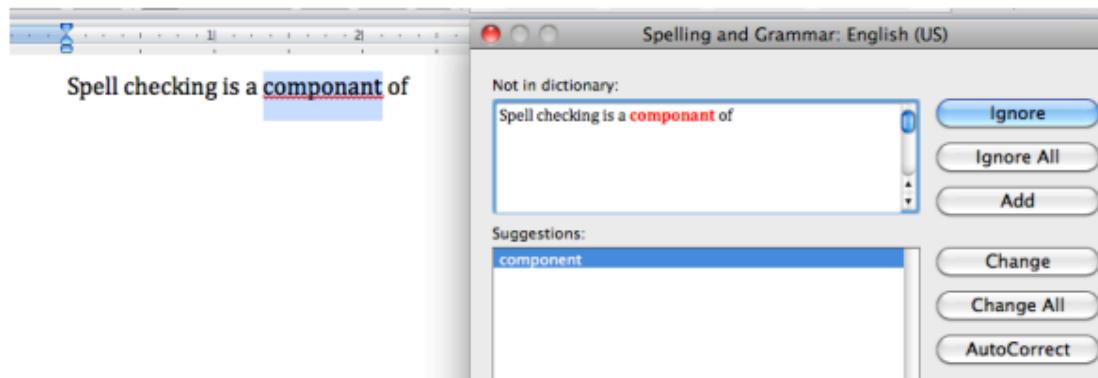
Test Data:

Spelling	Context
???	... said the longtime <i>aid/aide</i> to the Mayor of St. ...
???	... will squander the <i>aid/aide</i> it receives from the ...

Spell check

Applications for spelling correction

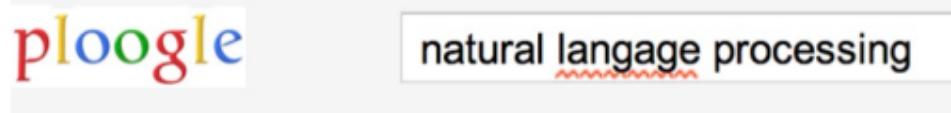
Word processing



Phones



Web search



Text Annotation Tasks

1. Classify the entire document
2. Classify individual word tokens
3. Identify phrases ("chunking")

Named Entity Recognition



CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by \$6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

NE Types



Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Type	Example
People	<i>Turing</i> is often considered to be the father of modern computer science.
Organization	The <i>IPCC</i> said it is likely that future tropical cyclones will become more intense.
Location	The <i>Mt. Sanitas</i> loop hike begins at the base of <i>Sunshine Canyon</i> .
Geo-Political Entity	<i>Palo Alto</i> is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the <i>Tappan Zee Bridge</i> or the <i>Lincoln Tunnel</i> .
Vehicles	The updated <i>Mini Cooper</i> retains its charm and agility.

Identifying phrases (chunking)

- Phrases that are useful for information extraction:
 - Named entities
 - As on previous slides
 - Relationship phrases
 - “said”, “according to”, ...
 - “was born in”, “hails from”, ...
 - “bought”, “hopes to acquire”, “formed a joint agreement with”, ...
 - Simple syntactic chunks (e.g., non-recursive NPs)
 - “Syntactic chunking” sometimes done before (or instead of) parsing
 - Also, “segmentation”: divide Chinese text into words (no spaces)
- So, how do we learn to mark phrases?
 - Earlier, we built an FST to mark dates by inserting brackets
 - But, it's common to set this up as a tagging problem ...

Reduce to a tagging problem ...

- The IOB encoding (Ramshaw & Marcus 1995):
 - B_X = “beginning” (first word of an X)
 - I_X = “inside” (non-first word of an X)
 - O = “outside” (not in any phrase)
 - Does not allow overlapping or recursive phrases

... **United Airlines** said Friday it has increased ...

B_ORG I_ORG O O O O O
... the move , spokesman **Tim Wagner** said ...
O O O O B_PER I_PER O



What if this were tagged as B_ORG instead?

Text Annotation Tasks

1. Classify the entire document
2. Classify individual word tokens
3. Identify phrases ("chunking")
4. Syntactic annotation (parsing)

Parser Evaluation Metrics

- Runtime
- Exact match
 - Is the parse 100% correct?
- Labeled precision, recall, F-measure of constituents
 - Precision: You predicted (NP,5,8); was it right?
 - Recall: (NP,5,8) was right; did you predict it?
- Easier versions:
 - Unlabeled: Don't worry about getting (NP,5,8) right, only (5,8)
 - Short sentences: Only test on sentences of ≤ 15 , ≤ 40 , ≤ 100 words
 - **Dependency parsing**: Labeled and unlabeled attachment accuracy
- Crossing brackets
 - You predicted (...5,8), but there was really a constituent (...6,10)

Labeled Dependency Parsing

Raw sentence

He reckons the current account deficit will narrow to only 1.8 billion in September.

POS-tagged sentence

He reckons the current account deficit will narrow to only 1.8 billion in September.

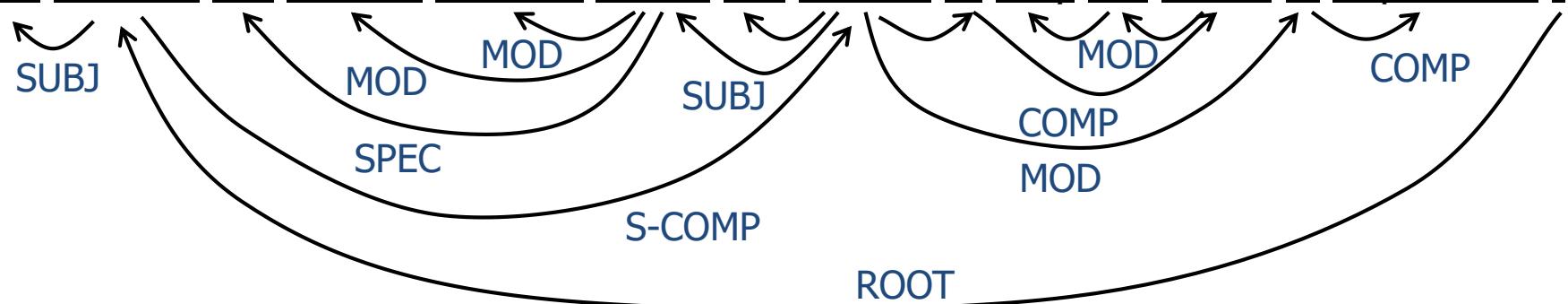
PRP VBZ DT JJ NN NN MD VB TO RB CD CD IN NNP .



Part-of-speech tagging

Word dependency parsed sentence

He reckons the current account deficit will narrow to only 1.8 billion in September .



Text Annotation Tasks

1. Classify the entire document
2. Classify individual word tokens
3. Identify phrases ("chunking")
4. Syntactic annotation (parsing)
5. Semantic annotation

- For each predicate (e.g., verb)
 1. find its arguments (e.g., NPs)
 2. determine their **semantic roles**

John drove Mary from Austin to Dallas in his Toyota Prius.

The hammer broke the window.

- **agent**: Actor of an action
- **patient**: Entity affected by the action
- **source**: Origin of the affected entity
- **destination**: Destination of the affected entity
- **instrument**: Tool used in performing action.
- **beneficiary**: Entity for whom action is performed

Might be helped by syntactic parsing first ...

- Consider one verb at a time: “bit”
- Classify the role (if any) of each of the 3 NPs

Color Code:

not-a-role

agent

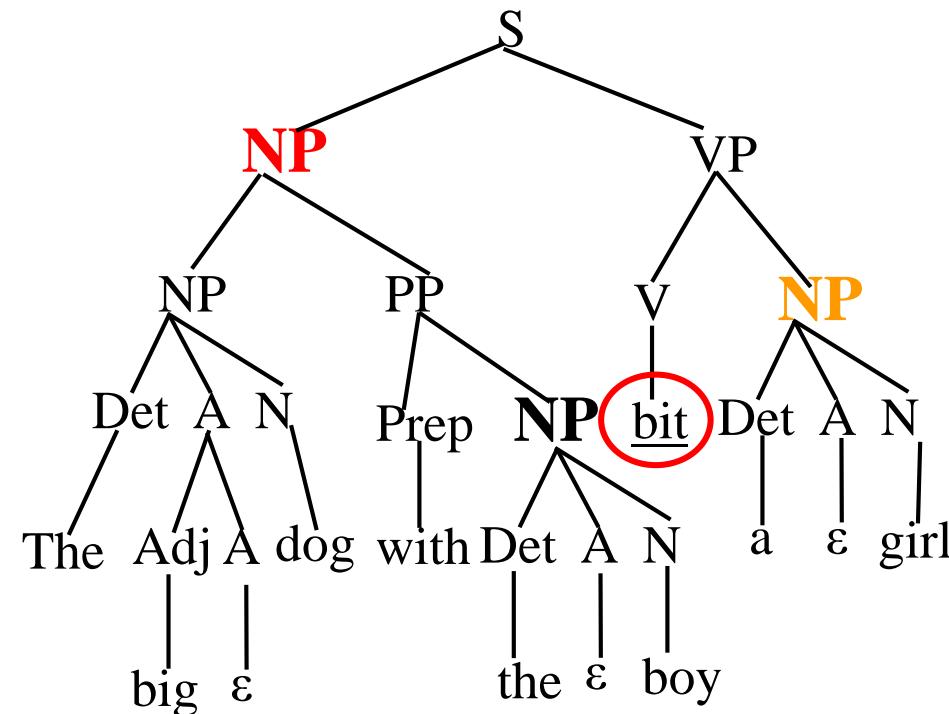
patient

source

destination

instrument

beneficiary

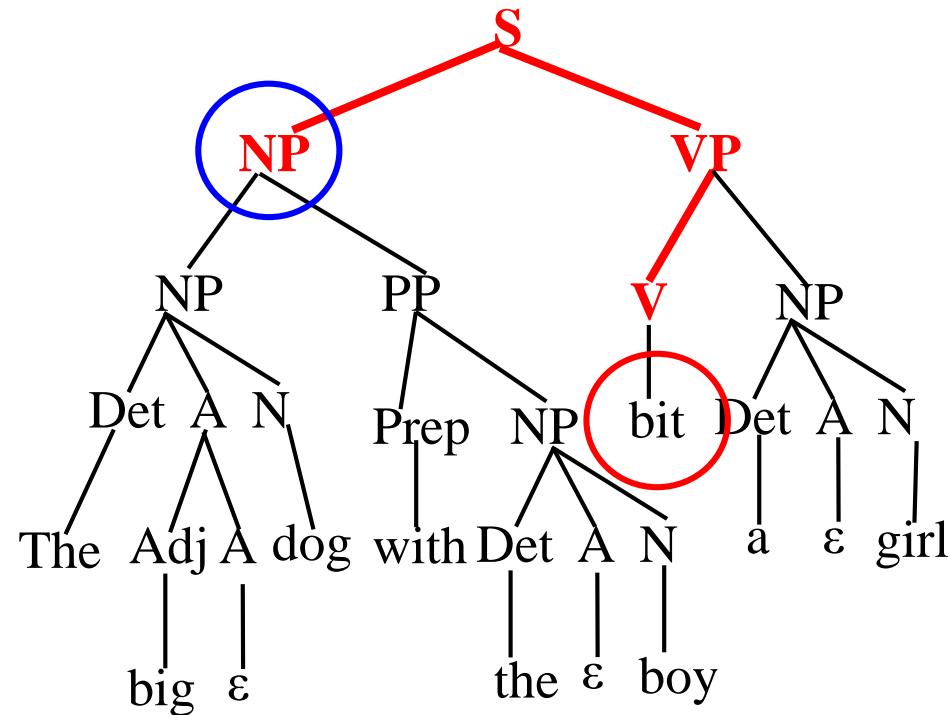


Parse tree paths as classification features

Path feature is

$V \uparrow VP \uparrow S \downarrow NP$

which tends to
be associated
with agent role



Head words as features

- Some roles prefer to be filled by certain kinds of NPs.
- This can give us useful features for classifying accurately:
 - “John ate the spaghetti with **chopsticks**.” (**instrument**)
“John ate the spaghetti with **meatballs**.” (**patient**)
“John ate the spaghetti with Mary.”
 - Instruments should be tools
 - Patient of “eat” should be edible
 - “John bought the car for **\$21K**.” (**instrument**)
“John bought the car for **Mary**.” (**beneficiary**)
 - Instrument of “buy” should be Money
 - Beneficiaries should be animate (things with desires)
 - “John drove Mary to school in the van”
“John drove the van to work with Mary.”
 - What do you think?

Semantic roles can feed into



further tasks

- Find the answer to a user's question
 - "Who" questions usually want Agents
 - "What" question usually want Patients
 - "How" and "with what" questions usually want Instruments
 - "Where" questions frequently want Sources/Destinations.
 - "For whom" questions usually want Beneficiaries
 - "To whom" questions usually want Destinations
- Generate text
 - Many languages have specific syntactic constructions that must or should be used for specific semantic roles.
- Word sense disambiguation, using selectional restrictions
 - The **bat** ate the **bug**. (what kind of bat? what kind of bug?)
 - Agents (particularly of "eat") should be animate – animal bat, not baseball bat
 - Patients of "eat" should be edible – animal bug, not software bug
 - John fired the secretary.
John **fired** the rifle.

- Remainder
 - Removing
 - Render_nonfunctional
 - Reparation
 - Reporting
 - Request
 - Reshaping
 - Residence
 - Rest
 - Revenge

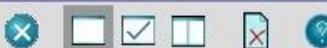
- Avenger <F1>
 - Injured_Party <>
 - Injury <F3>
 - Offender <F3>
 - Punishment <F12>
 - Degree <G>
 - Instrument <F3>
 - Manner <M>
 - Place <F3>
 - Time <F2>
 - + Depictive <D>
 - + Purpose <F4>
 - + Result <E>

- avenge.v
 - Lemma(V)
 - rcoll-brother [1/1]
 - rcoll-death [5/12]

- It will do no good t
 - With this, El Cid a
 - His secret ambition
 - For his distraught f
 - In Article 3 of the
 - The nausea threaten
 - Suddenly he walked
 - In Scaramouche the r
 - ^ Are you planning t
 - To avenge the death
 - The Trojans wish to
 - Did someone in this

- rcoll-defeat [5/16]
 - rcoll-father [0/3]
 - rcoll-murder [2/4]
 - np-ppagainst [0/1]
 - np-ppfor [1/2]
 - np-ppop [2/6]

SubCorpus Editor: V-429-s20-rcoll-death (77339)



0 It will do no good to AVENGE my death by killing him .

With this, El Cid at once AVENGED the death of his son and once again showed that any attempt to reconquer Valencia was fruitless while he still lived.

2 His secret ambition was for the Argentine ban to be lifted so he could get to England and AVENGE Pedro's death by taking out the English and especially one poker-faced Guards Officer. DNI

3 For his distraught family , only hanging would have AVENGED the death of the father of four

In Article 3 of the agreement, each had promised to AVENGE the violent death of the other with the blood of the murderer. DNI.

EE **GE** **PT** **Other** **Verb** **Sent**

Appositive	Comp <F3>	Ext <F1>	Gen <F5>
Head <F4>	Mod <F6>	Obj <F2>	Quant

Information Extraction



As a task:

Filling slots in a database from sub-segments of text.

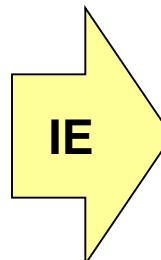
October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation CEO Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Phrase Types to Identify for IE



Closed set

U.S. states

He was born in Alabama...

The big Wyoming sky...

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

...was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

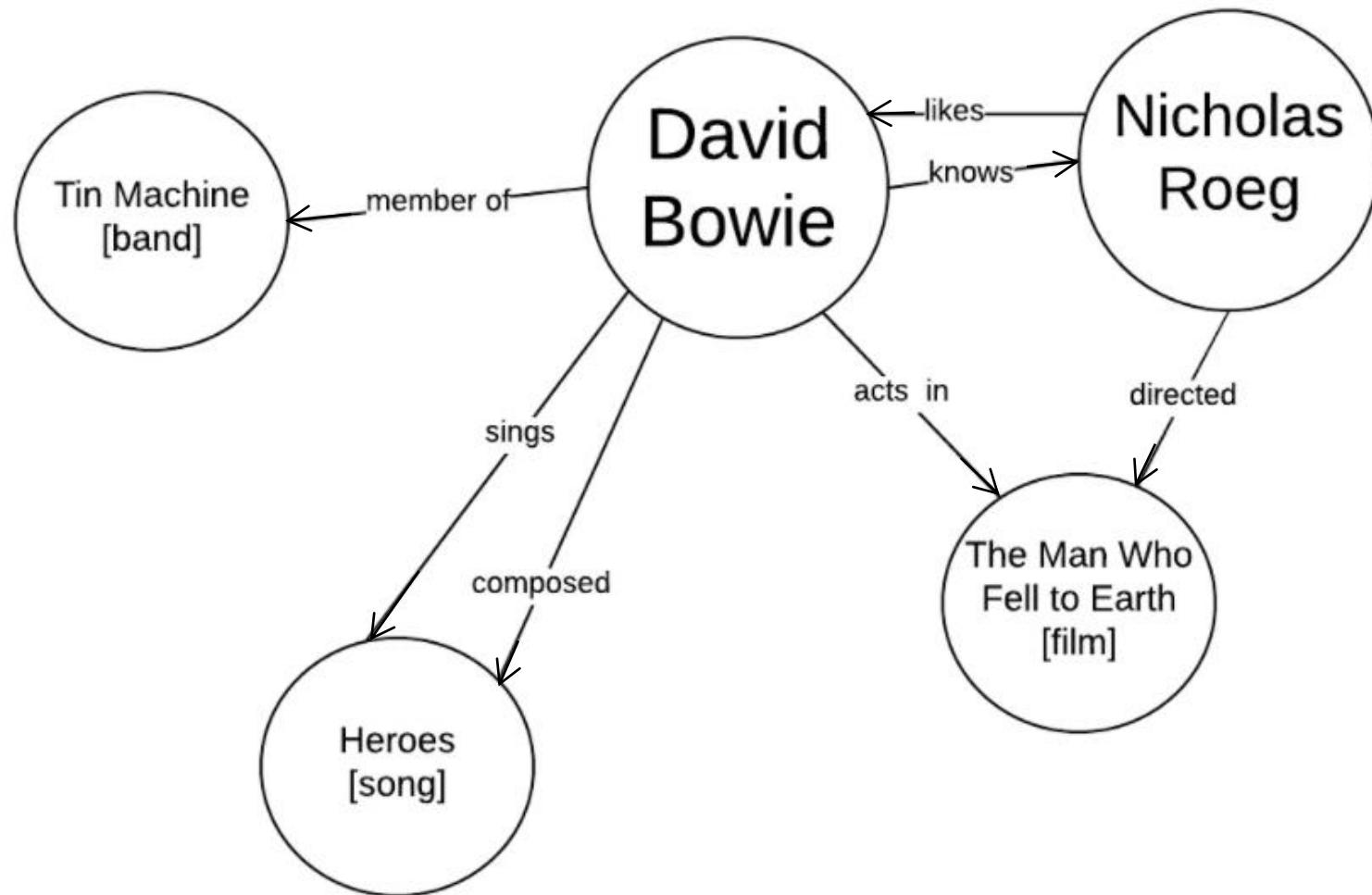
Example applications for IE

- Classified ads
- Restaurant reviews
- Bibliographic citations
- Appointment emails
- Legal opinions
- Papers describing clinical medical studies
- ...
- Adding facts to the semantic web

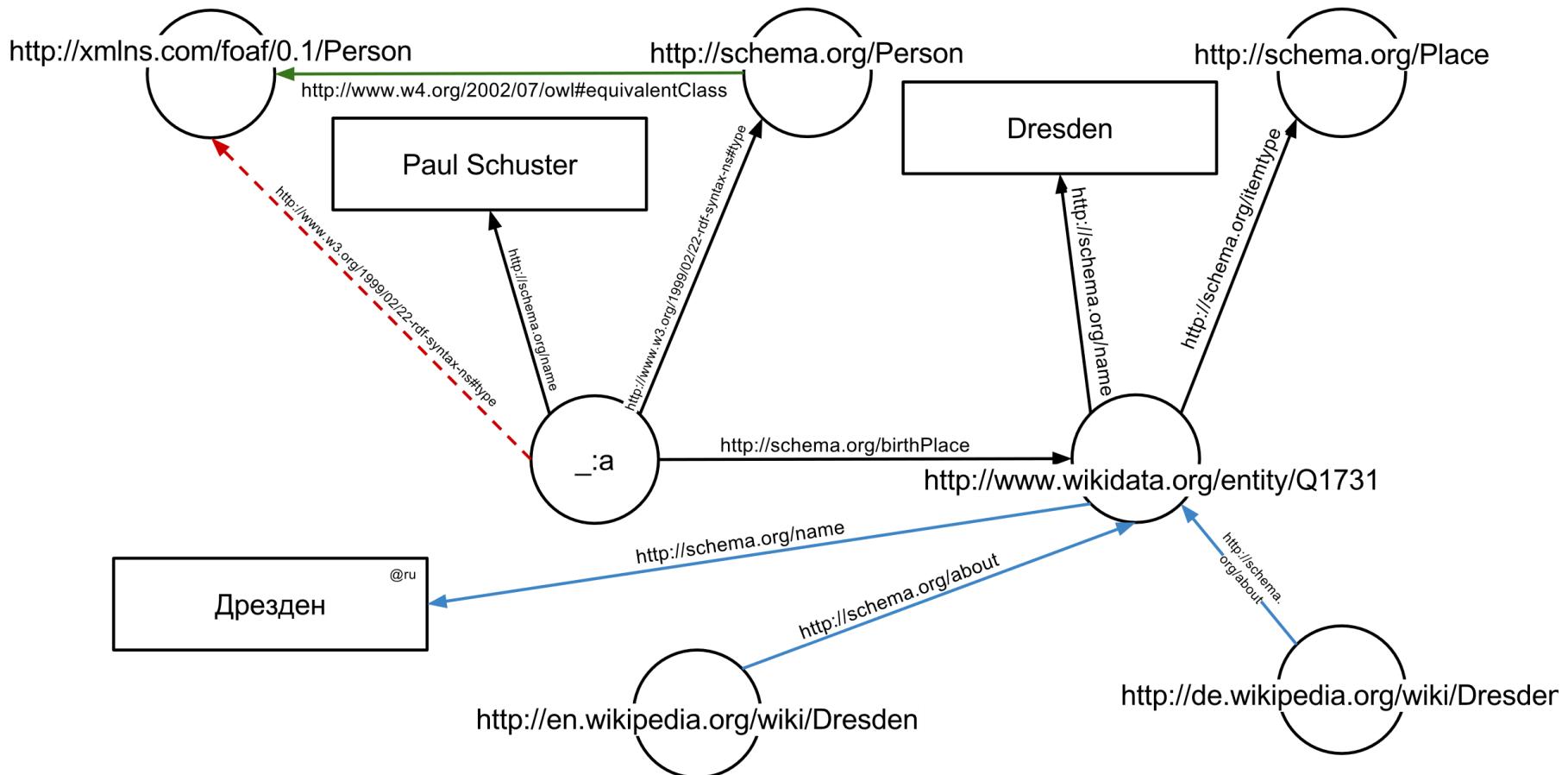
The Semantic Web

- A simple scheme for representing factual knowledge as a labeled graph
- Many information extraction tasks aim to produce something like this
- *[examples on next slides]*

The Semantic Web



The Semantic Web



The Semantic Web

- A simple scheme for representing factual knowledge as a labeled graph
- Many information extraction tasks aim to produce something like this
- Is a labeled graph (triples) really enough?
 - ☺ Can transform k-tuples to triples
(cf. Davidsonian event variable)
 - ☹ Supports facts about individuals, but no direct support for quantifiers or reasoning

Semantic Parsing

- Syntactic parsing maps from a sentence to a syntactic structure
- Semantic parsing maps from a sentence directly to a semantic structure
 - Often executable code, like a SQL query

Generating new text

1. Speech recognition (transcribe as text)
 2. Machine translation
 3. Text generation from semantics
 4. Inflect, analyze, pronounce, or transliterate words
 5. Single- or multi-doc summarization
-

Deeper Information Extraction

1. Coreference resolution (within a document)
2. Entity linking (across documents)
3. Event extraction and linking
4. Knowledge base population (KBP)

User interfaces

1. Dialogue systems

- Task-oriented dialogue (specific domains)
- Chatbots (unrestricted topic, fun or useful)
- Human-computer collaboration
- Interactive teaching

2. Language teaching; writing help

3. Question answering

4. Information retrieval

Multimodal interfaces or modeling

1. Sign languages
2. Speech + gestures
3. Images + captions
4. Brain recordings, human reaction times

Discovering Linguistic Structure

NLP automates things that humans do well, so that they can be done automatically on more sentences. But this slide is about language analysis that's hard even for humans. Computational linguistics (like comp bio, etc.) can discover underlying patterns in large datasets: things we didn't know! Some of these may already be hidden in vector representations ...

1. Decipherment (discovery of the meaning of texts written in ancient or obscure languages or scripts.)
2. Word meanings
3. Grammar induction (syntax, morphology)
4. Topic modeling

Some Big Questions

- Neural nets are fluent at generating text, but do they really represent and reason about the world the text describes? Are their answers consistent? Can they explain them?
- How can models learn effectively through interaction with the world or with human teachers?
- What kinds of linguistic biases should we build in, and how? Huge Transformer LMs with enormous training sets work well, but can we find architectures that generalize like humans from much smaller datasets? (Or is that just pretraining + few-shot or fine-tuning?)

The NLP Research Community



- ## Papers

- ACL Anthology has nearly everything, free!
 - Over 80,000 papers!
 - Free-text searchable
 - Great way to learn about current research on a topic
 - New search interfaces currently available in beta
 - » Find recent or highly cited work; follow citations
 - Used as a dataset by various projects
 - Analyzing the text of the papers (e.g., parsing it)
 - Extracting a graph of papers, authors, and institutions (Who wrote what? Who works where? What cites what?)
- Google Scholar to sort by citation count / track citations

The NLP Research Community



- Chitchat
 - [arXiv](#) papers
 - Twitter accounts
 - NLP researchers with active accounts
(grad students, profs, industry folks)
 - Official conference accounts
 - “NLP Highlights” podcast
 - “NLP News” newsletter

The NLP Research Community



- Institutions
 - Universities: Many have 2+ NLP faculty
 - Several “big players” with many faculty
 - Some of them also have good linguistics, cognitive science, machine learning, AI
 - Companies:
 - Old days: AT&T Bell Labs, IBM
 - Now: Microsoft Research, Google Brain/DeepMind, FAIR, Amazon, startups ...
 - Nonprofits: AI2, HuggingFace, TTIC, ...
 - Many niche markets – online reviews, medical transcription, news summarization, legal search and discovery ...

The NLP Research Community



- **Software**

- Lots of people distribute code for these tasks
 - Search github – fun to download packages and play around!
 - Or you can email a paper's authors to ask for their code
- Some lists of software, but no central site 😞
 - [PapersWithCode.com](https://paperswithcode.com)
 - Search for “awesome NLP” for some lists
- Toolkits and end-to-end pipelines for text analysis
 - [Hugging Face](https://huggingface.co) – > 20,000 models, > 2,000 datasets
 - Large pretrained models: pip install [transformers](#) ([quick tour](#))
 - Task-specific models: pip install [allennlp](#), etc.
 - [Allen NLP](#) (Python), [Spacy](#) (Cython), [UDPipe](#) (C++),
[Stanza](#) (Python), [CoreNLP](#) (Java), [NLTK](#) (Python)

The NLP Research Community



- **Datasets**

- Raw text or speech corpora
 - Or just their [n-gram counts](#), for super-big corpora
 - Various languages and genres
 - Usually there's some metadata (each document's date, author, etc.)
 - Sometimes \exists licensing restrictions (proprietary or copyright data)
- Text or speech with manual or automatic annotations
 - What kind of annotations? That's the rest of this lecture ...
 - May include translations into other languages
- Words and their relationships
 - [phonological](#), [morphological](#), [semantic](#), [translational](#), [evolutionary](#)
- [Grammars](#)
- [World Atlas of Linguistic Structures](#)
- Parameters of statistical models (e.g., grammar weights)

The NLP Research Community



• Datasets

- Read papers to find out what datasets others are using
 - [Linguistic Data Consortium](#) (searchable) hosts many large datasets
 - Many projects and competitions post data on their websites
 - But sometimes you have to email the author for a copy
- [CORPORA mailing list](#) is also good place to ask around
- [LREC Conference](#) publishes papers about new datasets & metrics
- Pay humans to annotate your data
 - Or to correct the annotations produced by your initial system
 - [Old task, new domain](#): Annotate parses etc. on *your* kind of data
 - [New task](#): Annotate something new that you want your system to find
 - [Auxiliary task](#): Annotate something new that your system may benefit from finding (e.g., annotate subjunctive mood to improve translation)
- Can you make annotation so much [fun](#) or so [worthwhile](#) that they'll do it for free?



Dr. Chetana is an Associate Professor in the CSIS department at Work Integrated Learning Programmes Division, BITS Pilani. She has more than 25 years of teaching and industry experience. She did her PhD in Computer Science and Engineering from a joint programme of IIT Bombay and Monash University, Australia. She has been working extensively on different state of art research projects and has been awarded the “Best Industry Aligned Research” at the CSI TechNext India 2019 - Awards to Academia. She has published various papers and is also a reviewer at national and international level peer reviewed conferences and journals. Her areas of expertise include Machine Learning, Natural Language Processing, Semantic Web, Deep Learning, Text Mining, Big Data Analytics, Information Retrieval and Software Engineering.

Thank you!!

Good References

<https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html>

<https://web.stanford.edu/~jurafsky/slp3/>

<https://www.geeksforgeeks.org/top-7-applications-of-natural-language-processing/>

<https://emerj.com/partner-content/nlp-current-applications-and-future-possibilities/>

<https://www.nltk.org/book/>

<https://www.coursera.org/learn/python-text-mining/home/week/1>

<https://openai.com/api/>

<https://analyticssteps.com/blogs/top-nlp-tools>

<https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html>

https://www.cstr.ed.ac.uk/emasters/course/natural_lang.html

<https://web.stanford.edu/class/cs224u/2016/materials/cs224u-2016-intro.pdf>

<https://www.mygreatlearning.com/blog/trending-natural-language-processing-applications/>

<https://grammarchecker.io/page/plagiarism-checker>

References

- Semantic Web stack of standards, W3C®
- Linking Open Data cloud diagram by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net>
- <https://www.youtube.com/watch?v=otH29Uoo-HE>
- <https://www.youtube.com/watch?v=UVfwBqcnbM>
- <https://www.youtube.com/watch?v=BELFUhssjXE>
- Overview of NLP Tasks and Featurization Frank Ferraro – ferraro@umbc.edu
- GEnerative AI for everyone- coursera
- Generative AI with LLM- coursera



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,

IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



**Session 2:
Date – 3rd Dec 2023
Time – 1.40 pm to 3.40 pm**

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Dan Jurafsky and many others who made their course materials freely available online.

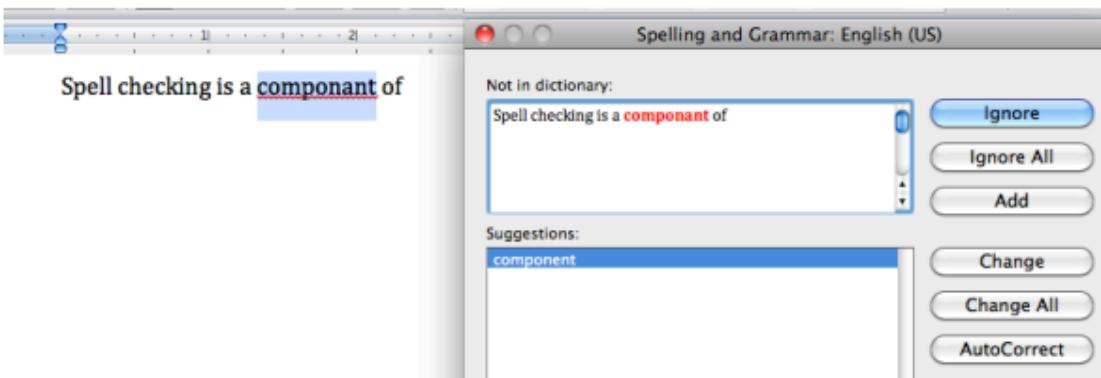
Session Content

- Spelling Correction and the Noisy Channel
 - Applications
 - Real World Spelling errors
 - Noisy channel
-

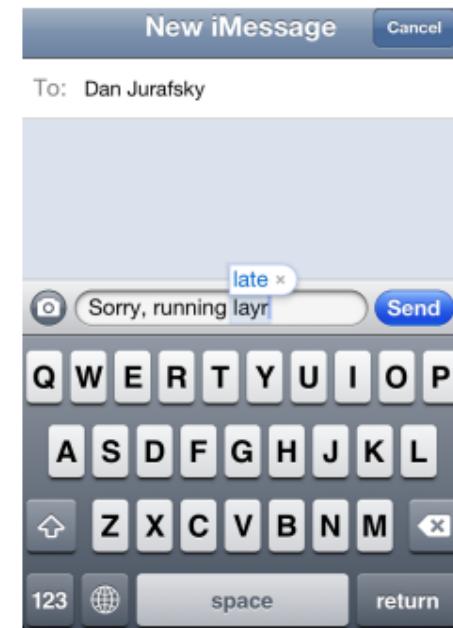
Spell check

Applications for spelling correction

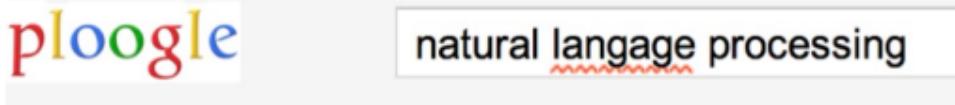
Word processing



Phones



Web search



Showing results for [natural language processing](#)
Search instead for [natural langage processing](#)

Spelling Error Detection

- Spelling Error Correction:
- Autocorrect
 - the->the
- Suggest a correction
- Suggestion lists

Types of spelling errors

- **Non-word Errors**
 - graffe -> giraffe
 - acress
- **Real-word Errors**
 - Typographical errors
 - three -> there
- **Cognitive Errors (homophones)**
 - piece -> peace
 - too -> two
- **Short forms/Slang:** In this case may be u r just being kewl.

Rates of spelling errors

- 26%: Web queries Wang et al. 2003
- 13%: Retyping, no backspace: Whitelaw et al. English German
- 7%: Words corrected retyping on phone-sized organizer
- 2%: Words uncorrected on organizer Soukoreff MacKenzie 2003
- 1-2%: Retyping: Kane and Wobbrock 2007, Gruden et al. 198

Non-word spelling error

- Non-word spelling error detection:
 - Any word not in a dictionary is an error
 - The larger the dictionary the better
- Non-word spelling error correction:
 - Generate candidates: real words that are similar to error
 - Choose the one which is best:
 - Shortest weighted edit distance
 - Highest noisy channel probability

Real word spelling errors

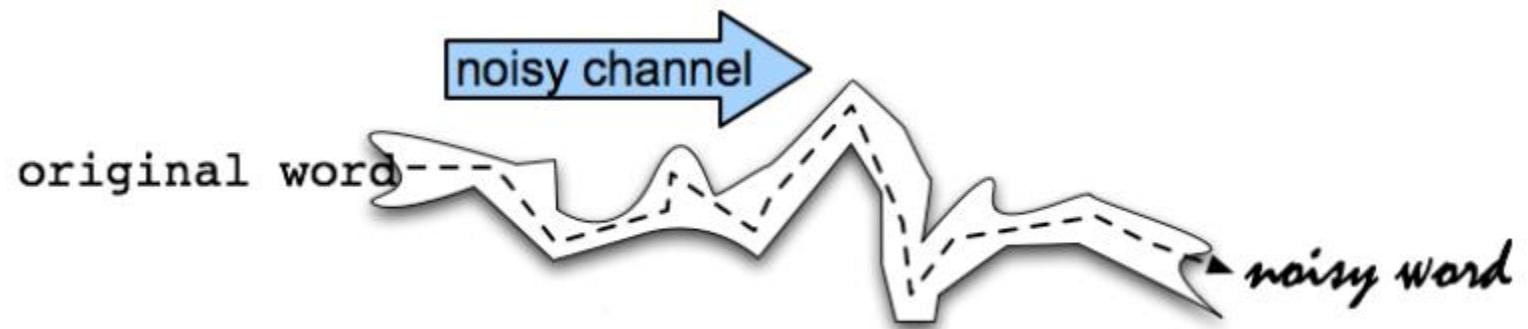
- For each word w , generate candidate set:
 - Find candidate words with similar pronunciations
 - Find candidate words with similar spelling
 - Include w in candidate set
- Choose best candidate
 - Noisy Channel
 - Classifier

Terminology

- These are character bigrams:
 - *st, pr, an ...*
- These are word bigrams:
 - *palo alto, flying from, road repairs*
- In today's class, we will generally deal with *word bigrams*
- In the accompanying Coursera lecture, we mostly deal with *character bigrams* (because we cover stuff complementary to what we're discussing here)

Similarly trigrams,
k-grams etc

Noisy Channel Intuition



Noisy Channel

We see an observation x of a misspelled word
Find the correct word w

$$\hat{w} = \operatorname{argmax}_{w \in V} P(w | x)$$

$$= \operatorname{argmax}_{w \in V} \frac{P(x | w)P(w)}{P(x)}$$

$$= \operatorname{argmax}_{w \in V} P(x | w)P(w)$$

Non-word spelling error example

acress

Candidate generation

- Words with similar spelling
 - Small edit distance to error
- Words with similar pronunciation
 - Small edit distance of pronunciation to error

Damerau-Levenshtein edit distance

- Minimal edit distance between two strings, where edits are:
 - Insertion
 - Deletion
 - Substitution
 - Transposition of two adjacent letters

Words within 1 of acress

Error	Candidate Correction	Correct Letter	Error Letter	Type
acress	actress	t	-	deletion
acress	cress	-	a	insertion
acress	caress	ca	ac	transposition
acress	access	c	r	substitution
acress	across	o	e	substitution
acress	acres	-	s	insertion
acress	acres	-	s	insertion

Candidate generation

- 80% of errors are within edit distance 1
- Almost all errors within edit distance 2
- Also allow insertion of space or hyphen
 - thisidea -> this idea
 - inlaw -> in-law

How do you generate the candidates?

1. Run through dictionary, check edit distance with each word
2. Generate all words within edit distance $\leq k$ (e.g., $k = 1$ or 2) and then intersect them with dictionary
3. Use a character k -gram index and find dictionary words that share “most” k -grams with word (e.g., by Jaccard coefficient)
4. Compute them fast with a Levenshtein finite state transducer
5. Have a precomputed map of words to possible corrections

A paradigm ...

- We want the best spell corrections
- Instead of finding the very best, we
 - Find a subset of pretty good corrections
 - (say, edit distance at most 2)
 - Find the best amongst them
- *These may not be the actual best*
- This is a recurring paradigm in IR including finding the best docs for a query, best answers, best ads ...
 - Find a good candidate set
 - Find the top K amongst them and return them as the best

Let's say we've generated candidates: Now back to Bayes' Rule

- We see an observation x of a misspelled word
- Find the correct word \hat{w}

$$\hat{w} = \operatorname{argmax}_{w \in V} P(w | x)$$

$$= \operatorname{argmax}_{w \in V} \frac{P(x | w)P(w)}{P(x)}$$

$$= \operatorname{argmax}_{w \in V} P(x | w)P(w)$$

What's $P(w)$?

Language Model

- Take a big supply of words (your document collection with T tokens); let $C(w)$ = # occurrences of w

$$P(w) = \frac{C(w)}{T}$$

Language Model

- Use any of the language modeling algorithms we've learned
 - Unigram, bigram, trigram
 - Web-scale spelling correction
 - Stupid backoff

Unigram Prior probability

- Counts from 404,253,213 words in Corpus of Contemporary English (COCA)

word	Frequency of word	P(word)
actress	9,321	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37,038	.0000916207
across	120,844	.0002989314
acres	12,874	.0000318463

Channel model probability

- Error model probability, Edit probability
 - Kernighan, Church, Gale 1990
- Misspelled word $x = x_1, x_2, x_3 \dots x_m$
- Correct word $w = w_1, w_2, w_3, \dots, w_n$
- $P(x|w)$ = probability of the edit
 - (deletion/insertion/substitution/transposition)

Computing error probability: confusion matrix

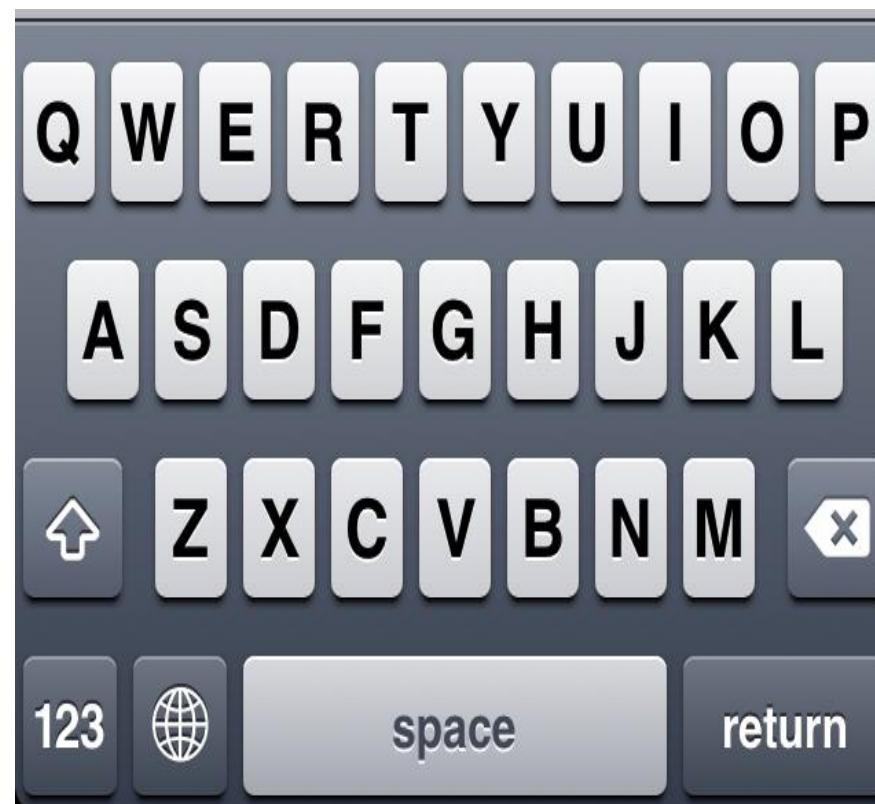
```
del[x,y]:    count(xy typed as x)
ins[x,y]:    count(x typed as xy)
sub[x,y]:    count(x typed as y)
trans[x,y]:  count(xy typed as yx)
```

Insertion and deletion conditioned on previous character

Confusion matrix for spelling errors

X	sub[X, Y] = Substitution of X (incorrect) for Y (correct)																									
	Y (correct)																									
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	0	8	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Nearby keys



Generating the confusion matrix

- Peter Norvig's list of errors
- Peter Norvig's list of counts of single-edit errors
 - All Peter Norvig's ngrams data links: <http://norvig.com/ngrams/>

Channel Model

$$P(x|w) = \begin{cases} \frac{\text{del}[w_{i-1}, w_i]}{\text{count}[w_{i-1} w_i]}, & \text{if deletion} \\ \frac{\text{ins}[w_{i-1}, x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\ \frac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, & \text{if substitution} \\ \frac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_i w_{i+1}]}, & \text{if transposition} \end{cases}$$

Smoothing probabilities: Add-1 smoothing

- But if we use the confusion matrix example, unseen errors are impossible!
- They'll make the overall probability 0. That seems too harsh
 - e.g., in Kernighan's chart $q \rightarrow a$ and $a \rightarrow q$ are both 0, even though they're adjacent on the keyboard!
- A simple solution is to add 1 to all counts and then if there is a $|A|$ character alphabet, to normalize appropriately:

$$\text{If substitution, } P(x | w) = \frac{\text{sub}[x, w] + 1}{\text{count}[w] + A}$$

Channel model for acres

Candidate Correction	Correct Letter	Error Letter	x w	P(x word)
actress	t	-	c ct	.000117
cress	-	a	a #	.00000144
caress	ca	ac	ac ca	.00000164
access	c	r	r c	.000000209
across	o	e	e o	.0000093
acres	-	s	es e	.0000321
acres	-	s	ss s	.0000342

Noisy channel probability for acres

Candidate Correction	Correct Letter	Error Letter	x w	P(x word)	P(word)	$10^9 * P(x w)P(w)$
actress	t	-	c ct	.000117	.0000231	2.7
cress	-	a	a #	.00000144	.000000544	.00078
caress	ca	ac	ac ca	.00000164	.00000170	.0028
access	c	r	r c	.000000209	.0000916	.019
across	o	e	e o	.0000093	.000299	2.8
acres	-	s	es e	.0000321	.0000318	1.0
acres	-	s	ss s	.0000342	.0000318	1.0

Noisy channel probability for acres

Candidate Correction	Correct Letter	Error Letter	x w	P(x word)	P(word)	$10^9 * P(x w)P(w)$
actress	t	-	c ct	.000117	.0000231	2.7
cress	-	a	a #	.00000144	.000000544	.00078
caress	ca	ac	ac ca	.00000164	.00000170	.0028
access	c	r	r c	.000000209	.0000916	.019
across	o	e	e o	.0000093	.000299	2.8
acres	-	s	es e	.0000321	.0000318	1.0
acres	-	s	ss s	.0000342	.0000318	1.0

Incorporating context words: Context-sensitive spelling correction

- Determining whether **actress** or **across** is appropriate will require looking at the context of use
- We can do this with a better **language model**
- A **bigram language model** conditions the probability of a word on (just) the previous word

$$P(w_1 \dots w_n) = P(w_1)P(w_2 | w_1) \dots P(w_n | w_{n-1})$$

Incorporating context words

- For unigram counts, $P(w)$ is always non-zero
 - if our dictionary is derived from the document collection
- This won't be true of $P(w_k | w_{k-1})$. We need to **smooth**
- We could use add-1 smoothing on this conditional distribution
- But here's a better way – interpolate a unigram and a bigram:

$$P_{\text{li}}(w_k | w_{k-1}) = \lambda P_{\text{uni}}(w_k) + (1-\lambda)P_{\text{bi}}(w_k | w_{k-1})$$

- $P_{\text{bi}}(w_k | w_{k-1}) = C(w_{k-1}, w_k) / C(w_{k-1})$

All the important fine points

- Note that we have several probability distributions for words
 - Keep them straight!
- You might want/need to work with log probabilities:
 - $\log P(w_1 \dots w_n) = \log P(w_1) + \log P(w_2 | w_1) + \dots + \log P(w_n | w_{n-1})$
 - Otherwise, be very careful about floating point underflow
- Our query may be words anywhere in a document
 - We'll start the bigram estimate of a sequence with a unigram estimate
 - Often, people instead condition on a start-of-sequence symbol, but not good here
 - Because of this, the unigram and bigram counts have different totals – not a problem

Using a bigram language model

- “*a stellar and versatile **acress** whose combination of sass and glamour...*”
- Counts from the Corpus of Contemporary American English with add-1 smoothing
 - $P(\text{actress}|\text{versatile}) = .000021$ $P(\text{whose}|\text{actress}) = .0010$
 - $P(\text{across}|\text{versatile}) = .000021$ $P(\text{whose}|\text{across}) = .000006$
- $P(\text{"versatile actress whose"}) = .000021 * .0010 = 210 \times 10^{-10}$
- $P(\text{"versatile across whose"}) = .000021 * .000006 = 1 \times 10^{-10}$

Using a bigram language model

- “**a stellar and versatile **acress** whose combination of sass and glamour...**”
- Counts from the Corpus of Contemporary American English with add-1 smoothing
 - $P(\text{actress}|\text{versatile}) = .000021$ $P(\text{whose}|\text{actress}) = .0010$
 - $P(\text{across}|\text{versatile}) = .000021$ $P(\text{whose}|\text{across}) = .000006$
- $P(\text{"versatile actress whose"}) = .000021 * .0010 = 210 \times 10^{-10}$
- $P(\text{"versatile across whose"}) = .000021 * .000006 = 1 \times 10^{-10}$

Evaluation

- Some spelling error test sets
 - Wikipedia's list of common English misspelling
 - Aspell filtered version of that list
 - Birkbeck spelling error corpus
 - Peter Norvig's list of errors (includes Wikipedia and Birkbeck, for training or testing)

Real world spelling errors

- ...leaving in about fifteen **minuets** to go to her house.
- The design **an** construction of the system...
- Can they **lave** him my messages?
- The study was conducted mainly **be** John Black.
- 25-40% of spelling errors are real words Kukich 1992

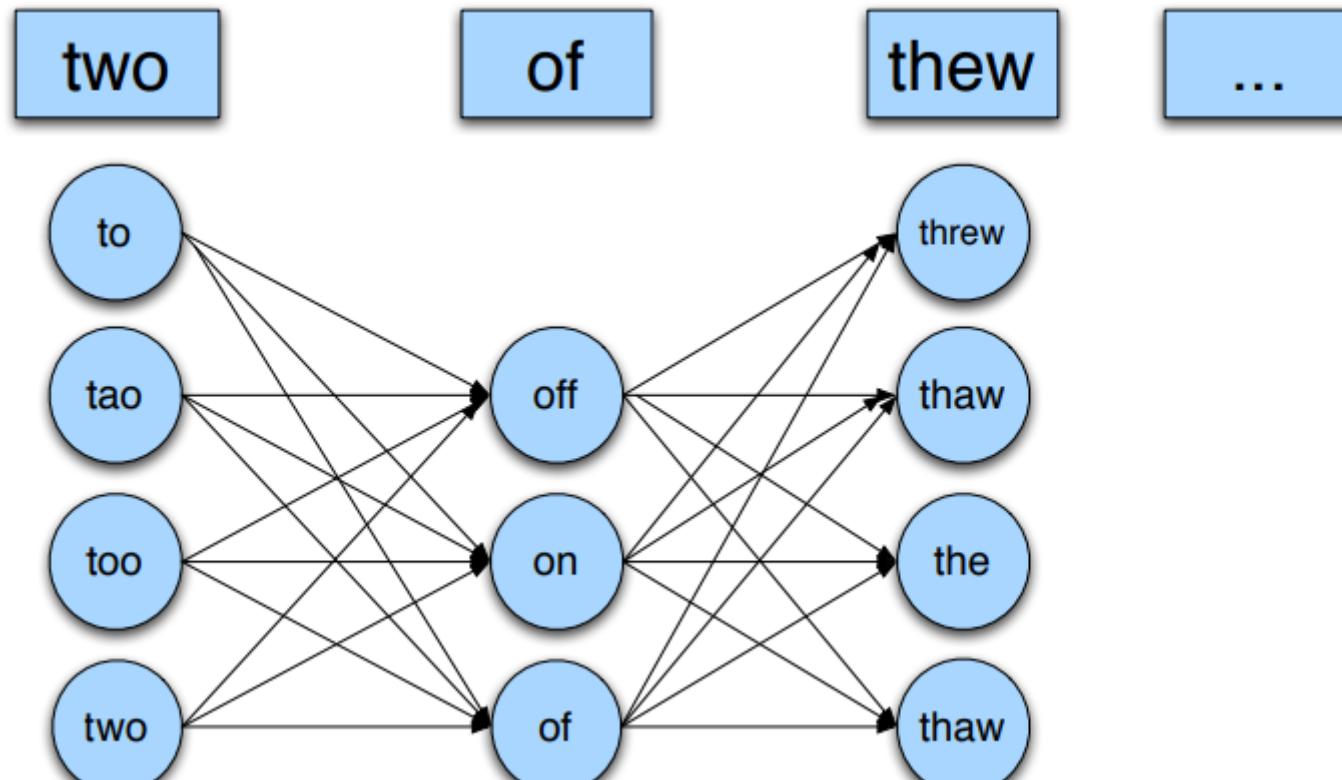
Solving real-world spelling errors

- For each word in sentence
 - Generate candidate set
 - the word itself
 - all single-letter edits that are English words
 - words that are homophones
 - Choose best candidates
 - Noisy channel model
 - Task-specific classifier

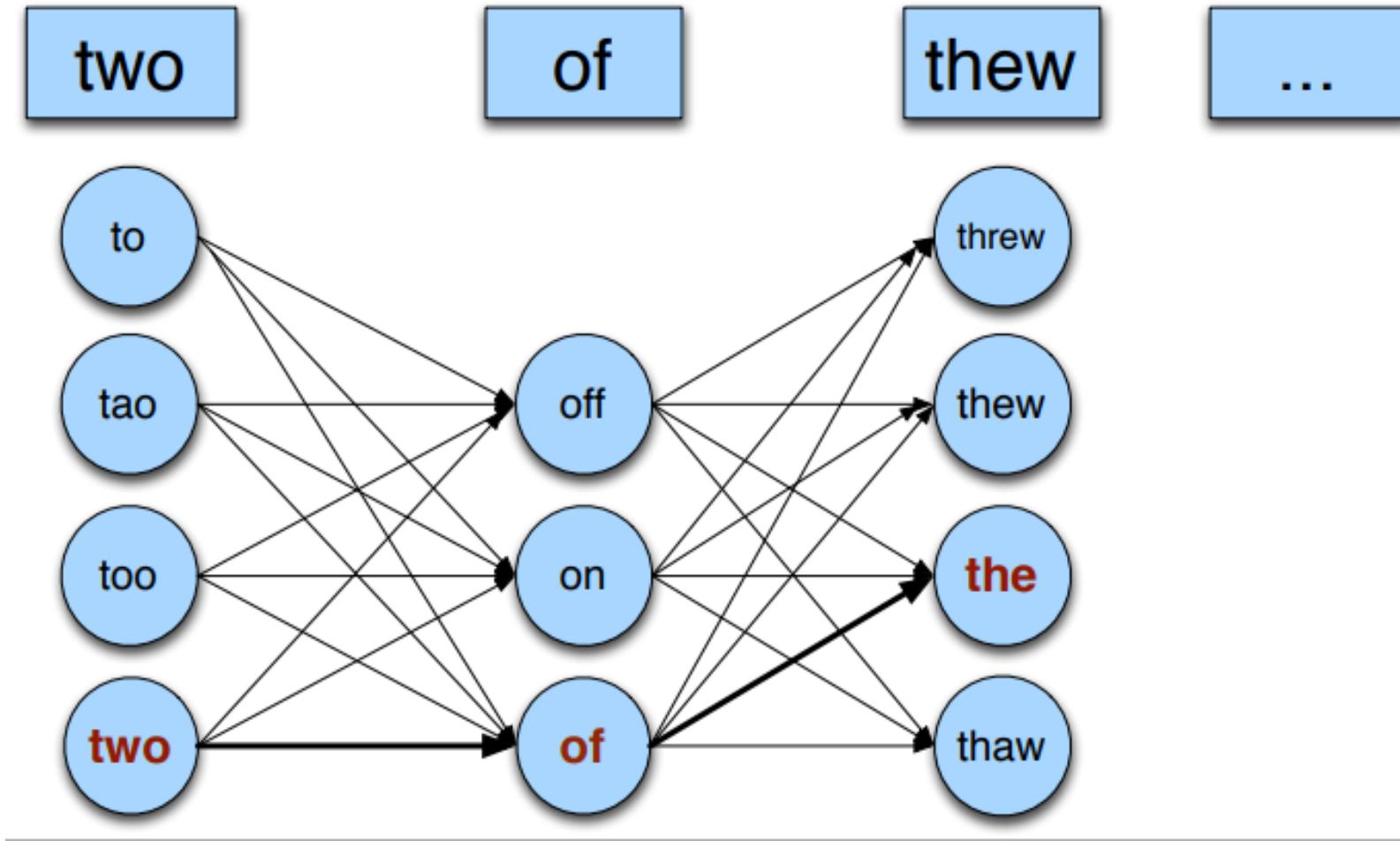
Noisy channel for real-word spell correction

- Given a sentence $w_1, w_2, w_3, \dots, w_n$
- Generate a set of candidates for each word w_i
 - Candidate(w_1) = $\{w_1, w'_1, w''_1, w'''_1, \dots\}$
 - Candidate(w_2) = $\{w_2, w'_2, w''_2, w'''_2, \dots\}$
 - Candidate(w_n) = $\{w_n, w'_n, w''_n, w'''_n, \dots\}$
- Choose the sequence W that maximizes $P(W)$

Noisy channel for real-word spell correction



Noisy channel for real-word spell correction



Simplification: One error per sentence

- Out of all possible sentences with one word replaced
 - w_1, w''_2, w_3, w_4 two off thew
 - w_1, w_2, w'_3, w_4 two of the
 - w'''_1, w_2, w_3, w_4 too of thew
 - ...
- Choose the sequence W that maximizes $P(W)$

Where to get the probabilities

- Language model
 - Unigram
 - Bigram Etc
- Channel model
 - Same as for non-word spelling correction
 - Plus need probability for no error, $P(w|w)$

Probability of no error

- What is the channel probability for a correctly typed word?
 - $P(\text{"the"} | \text{"the"})$
- Obviously this depends on the application
 - .90 (1 error in 10 words)
 - .95 (1 error in 20 words)
 - .99 (1 error in 100 words)
 - .995 (1 error in 200 words)

Peter Norvig's “thew” example

x	w	x w	P(x w)	P(w)	$10^9 P(x w)P(w)$
thew	the	ew e	0.000007	0.02	144
thew	thew		0.95	0.00000009	90
thew	thaw	e a	0.001	0.0000007	0.7
thew	threw	h hr	0.000008	0.000004	0.03
thew	thwe	ew we	0.000003	0.00000004	0.0001

HCI issues in spelling

- If very confident in correction
 - Autocorrect
- Less confident
 - Give the best correction
- Less confident
 - Give a correction list
- Unconfident
 - Just flag as an error

Improvements to channel model

- Allow richer edits (Brill and Moore 2000)
 - Ent -> ant
 - ph -> f
 - le -> al
- Incorporate pronunciation into channel (Toutanova and Moore 2002

Channel model

- Factors that could influence $p(\text{misspelling} \mid \text{word})$
 - The source letter
 - The target letter
 - Surrounding letters
 - The position in the word
 - Nearby keys on the keyboard
 - Homology on the keyboard
 - Pronunciations
 - Likely morpheme transformations

Classifier-based methods for real-word spelling correction

- Instead of just channel model and language model
- Use many features in a classifier
- Build a classifier for a specific pair like: whether/weather
 - “cloudy” within +- 10 words
 - ____ to VERB
 - ____ or not

Rule-based methods

- Rule-based methods rely on predefined rules and dictionaries to identify and correct spelling errors.
- For example, a rule-based spell checker might use a list of common words, a set of spelling rules, and a set of exceptions to detect and suggest corrections.
- Rule-based methods are easy to implement, fast, and transparent.
- However, they also have some limitations.
 - They cannot handle new words, slang, proper nouns, or variations in spelling.
 - They also require a lot of manual work to create and maintain the rules and dictionaries.
 - Moreover, they may not be able to handle complex or ambiguous errors, such as homophones, context-dependent errors, or errors that affect the meaning of the sentence.

Statistical methods

- Statistical methods use probabilistic models and machine learning to learn from large corpora of text and generate corrections based on the likelihood of occurrence.
- For example, a statistical spell checker might use a language model to estimate the probability of a word given its previous and next words, and a noisy channel model to estimate the probability of a word given its misspelled version.
- Statistical methods are more flexible, adaptive, and scalable than rule-based methods.
- They can handle new words, variations in spelling, and different domains and languages.
- However, they also have some drawbacks.
 - They require a lot of data and computational resources to train and run the models.
 - They may also produce inaccurate or inappropriate corrections, especially if the data is noisy, biased, or sparse.
 - Furthermore, they may not be able to explain why they made a certain correction or how confident they are about it.

Neural methods

- Neural methods use deep neural networks and artificial intelligence to learn from large corpora of text and generate corrections based on the semantic and syntactic features of the text.
- For example, a neural spell checker might use an encoder-decoder architecture to encode the misspelled word and its context into a vector representation, and then decode it into a corrected word or sequence.
- Neural methods are more powerful, robust, and generalizable than rule-based and statistical methods.
- They can handle complex and multiple errors, capture the meaning and structure of the text, and adapt to different domains and languages.
- However, they also have some challenges.
 - They require a lot of data and computational resources to train and run the networks.
 - They may also suffer from overfitting, underfitting, or hallucination, meaning that they may memorize, ignore, or invent words that are not in the data.
 - Moreover, they may not be able to explain why they made a certain correction or how confident they are about it.

NN algorithms used in spell check and grammar check

- **Sequence-to-Sequence (Seq2Seq) Models:**
- Based on recurrent neural networks (RNNs) or transformer architectures,
- These models are trained on pairs of input and output sequences (e.g., correct sentences and their corresponding erroneous versions) and learn to predict corrections by generating output sequences given the input.
- **LSTM (Long Short-Term Memory) Networks:**
- type of recurrent neural network, is effective in learning long-range dependencies in sequential data, making it useful for spellchecking and grammar correction tasks.
- LSTMs can handle variable-length sequences and retain information over longer periods, which is beneficial in understanding context within sentences.

Bi-directional LSTM

- *The bidirectional Long Short Term Memory networks (BLSTM) that can be trained using all available input information in the past and future of a specific time frame.*
- *For example,* Let us take an example of missing word generation in the *I am ___ student.*
- Unidirectional LSTMs will use only ‘I am’ to generate next word and based on the example it has seen during training it will generate a new word (it may be ‘a’, ‘very’ etc.).
- But bidirectional LSTMs have information of the past (I am) and future (student), so it can easily see that here it has to be **a**.

NN algorithms used in spell check and grammar check

- **Language Models for Contextual Understanding:**
- Language models trained on large corpora of text data, such as GPT-3, are adept at understanding context and can be fine-tuned for grammar and spellcheck tasks.
- These models can generate corrections based on the context provided in a given sentence.
- **Ensemble Models:**
- Combining multiple neural network models or ensemble techniques can enhance the overall accuracy of spellcheck and grammar correction systems.
- Ensemble methods leverage the strengths of various models to improve performance.

NN algorithms used in spell check and grammar check

- **Attention Mechanisms:**
- Commonly integrated into transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), or RoBERTa, enable the model to focus on specific parts of the input sequence while making corrections.
- This attention mechanism enhances the model's ability to capture relevant context for accurate spell and grammar checks.
- **BERT-Based Models:**
- BERT, a transformer-based language representation model, has been fine-tuned for various natural language understanding tasks, including grammar and spellcheck.
- Its bidirectional architecture allows it to understand context more effectively, aiding in better error detection and correction.

Combining Neural network and statistical methods

- While neural networks excel at learning complex patterns and representations from data, statistical methods can complement these networks by providing additional insights or features that improve the overall spellcheck process.
- Fusion of methodologies aims to improve the overall performance, reliability, and contextual understanding of the spellcheck systems, leading to more effective error detection and correction in natural language text

Combining Neural network and statistical methods

- **Language Models as Prior Knowledge:**
 - Statistical language models, such as n-grams or Hidden Markov Models (HMMs), can provide probabilities of word sequences or character sequences, which neural networks can leverage to refine their predictions or to provide additional context.
- **Error Probability Estimation:**
 - Statistical methods can help estimate the likelihood of a word being misspelled or a sequence of words containing grammatical errors.
 - These probabilities can serve as features or signals for the neural network model to consider during its learning process, enabling it to weigh the predictions accordingly.

Combining Neural network and statistical methods

- **Feature Engineering:**
 - Features derived from statistical analyses, such as character n-grams, syntactic features, or probabilities based on language rules, can be fed into the neural network as additional input features, enhancing its ability to detect and correct spelling or grammar errors.
- **Combining Outputs:**
 - ensemble the outputs of multiple neural network models.
 - For instance, by leveraging statistical techniques like voting or weighted averaging of predictions from different neural network architectures, the overall accuracy and robustness of the spellcheck system can be improved.

Combining Neural network and statistical methods

- **Scoring and Ranking:**
 - After obtaining predictions from a neural network model, statistical techniques can be used to rank potential corrections or errors based on their probabilities or scores, aiding in selecting the most likely and contextually appropriate corrections.

Combining Rule based, Statistical and Neural network methods

- **Rule-based Preprocessing:**
- Utilize rule-based systems to perform initial preprocessing steps such as basic spelling checks, identifying common grammatical errors, and applying language-specific rules.
- Create rules to detect and correct known errors or patterns that are easily identifiable, such as basic misspellings or common grammatical mistakes.

Combining Rule based, Statistical and Neural network methods

- **Statistical Methods for Error Detection:**
- Employ statistical models, such as n-grams, Hidden Markov Models (HMMs), or language models, to calculate probabilities of word sequences or character sequences.
- Use statistical analysis to identify anomalies or deviations in text that might indicate spelling or grammar errors.
- Estimate error probabilities for words or sequences and use these probabilities as features in the subsequent stages.

Combining Rule based, Statistical and Neural network methods

- **Neural Network Integration:**
- Train neural network architectures, such as LSTM networks, Transformers (e.g., BERT), or sequence-to-sequence models, to learn from labeled data containing correct and incorrect text pairs.
- Use neural networks to capture complex patterns, context, and semantic understanding in the text data.
- Incorporate statistical features or error probabilities derived from the statistical methods as additional input features for the neural network models.

Combining Rule based, Statistical and Neural network methods

- **Hybrid Model Ensemble:**
- Ensemble different models (rule-based, statistical, neural network) to combine their outputs for final error detection and correction.
- Develop a mechanism to combine predictions from each approach, either through voting mechanisms, weighted averaging, or stacking techniques.
- Apply a post-processing step that merges the strengths of each model's predictions to provide the most accurate and contextually appropriate corrections.

References

- Speech and Language processing: An introduction to Natural Language Processing, Computational Linguistics and speech Recognition by Daniel Jurafsky and James H. Martin[3rd edition] – Appendix B

Dan Jurafsky Course :

- https://www.youtube.com/watch?v=dAf_vcs7DuE

Pawan Goyal Course:

- <https://www.youtube.com/watch?v=Hmc0lVdXVpE>
- <https://sanketp.medium.com/language-models-spellchecking-and-autocorrection-dd10f739443c>
- <https://www.youtube.com/watch?v=4htP-hmlBvc>
- <https://nlp.stanford.edu/IR-book/html/htmledition/spelling-correction-1.html>
- <https://www.geeksforgeeks.org/spelling-checker-in-python/>
- <https://www.slideshare.net/AmirShokri2/personalized-spell-checking-using-neural-networks>
- https://www.youtube.com/watch?v=2RU5egfeb_g

References

- <https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html>
- <https://web.stanford.edu/~jurafsky/slp3/>
- <https://www.youtube.com/watch?v=3a7QEkrDko>
- <https://www.youtube.com/watch?v=sQVpMzNXdco>
- <https://towardsdatascience.com/textblob-spelling-correction-46321fc7f8b8>
- DL models for Spellcheck
- <https://bhashkarkunal.medium.com/spelling-correction-using-deep-learning-how-bi-directional-lstm-with-attention-flow-works-in-366fabcc7a2f#:~:text=In%20Symmetric%20delete%20spelling%20correction,during%20a%20pre%2Dcalculation%20step.>



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,

IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



**Session 2:
Date – 10th December 2023
Time – 1.40pm to 3.40 pm**

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Dan Jurafsky and many others who made their course materials freely available online.

Session Content

-
- Grammatical Errors
 - Types of Grammatical Errors
 - Corpora and Tasks
 - Methodologies and Systems
 - Grammar Correction Approaches
 - Evaluation
 - Current and Future Trends
-

What is a grammatical error?

Syntax:

“Each language has its own systematic ways through which words and sentences are assembled to convey meaning.” Fraser & Hodson (1978)

Syntax errors are rule-driven (e.g. subj-verb agreement) thus easier to learn

Usage: Conventional usage habits

A wrong preposition or missing determiner

– do not break rules of syntax but of usage.

Usage errors are most common for learners

– greater reliance on memory than rules

Syntax

- **Syntax** (of natural languages) describe how words are strung together to form components of sentences, and how those components are strung together to form sentences.
- In the core of the description of the syntax of a natural language, we use context-free grammars (CFGs).
- Groups of words may behave as a single unit or phrases, called as **constituent**.
 - noun phrase,
 - verb phrase
- CFGs will allow us to model these constituency facts.

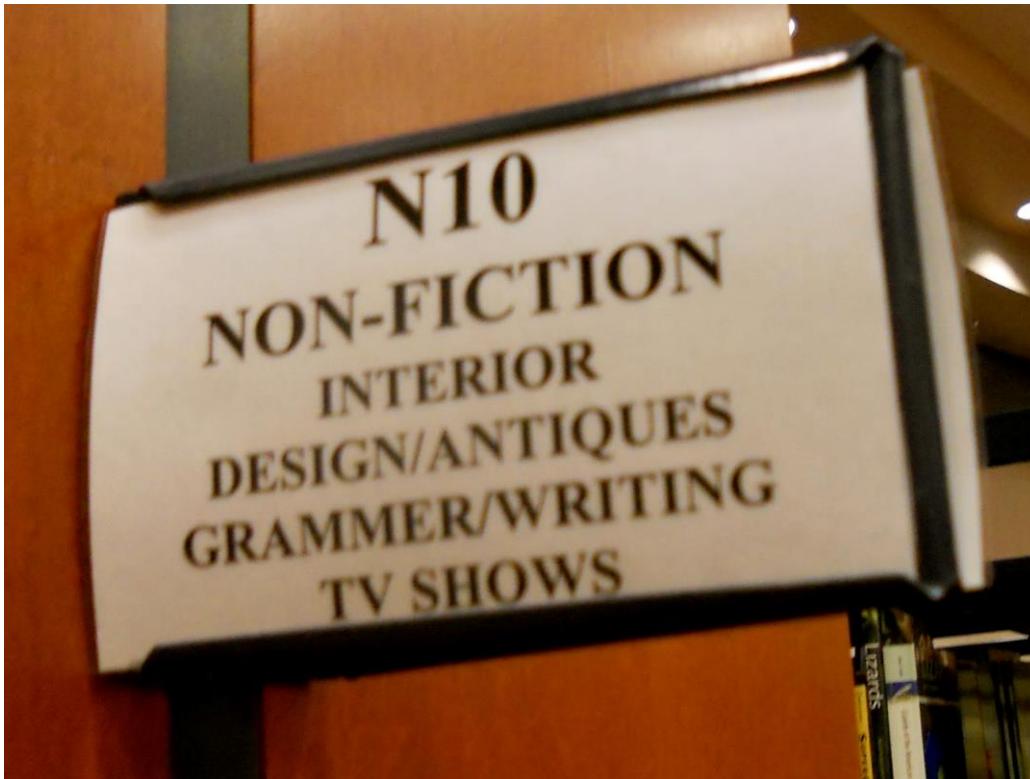
Applications of Grammar Check

- Question Answering
- Information Extraction
- Generation
- Translation
- Understanding

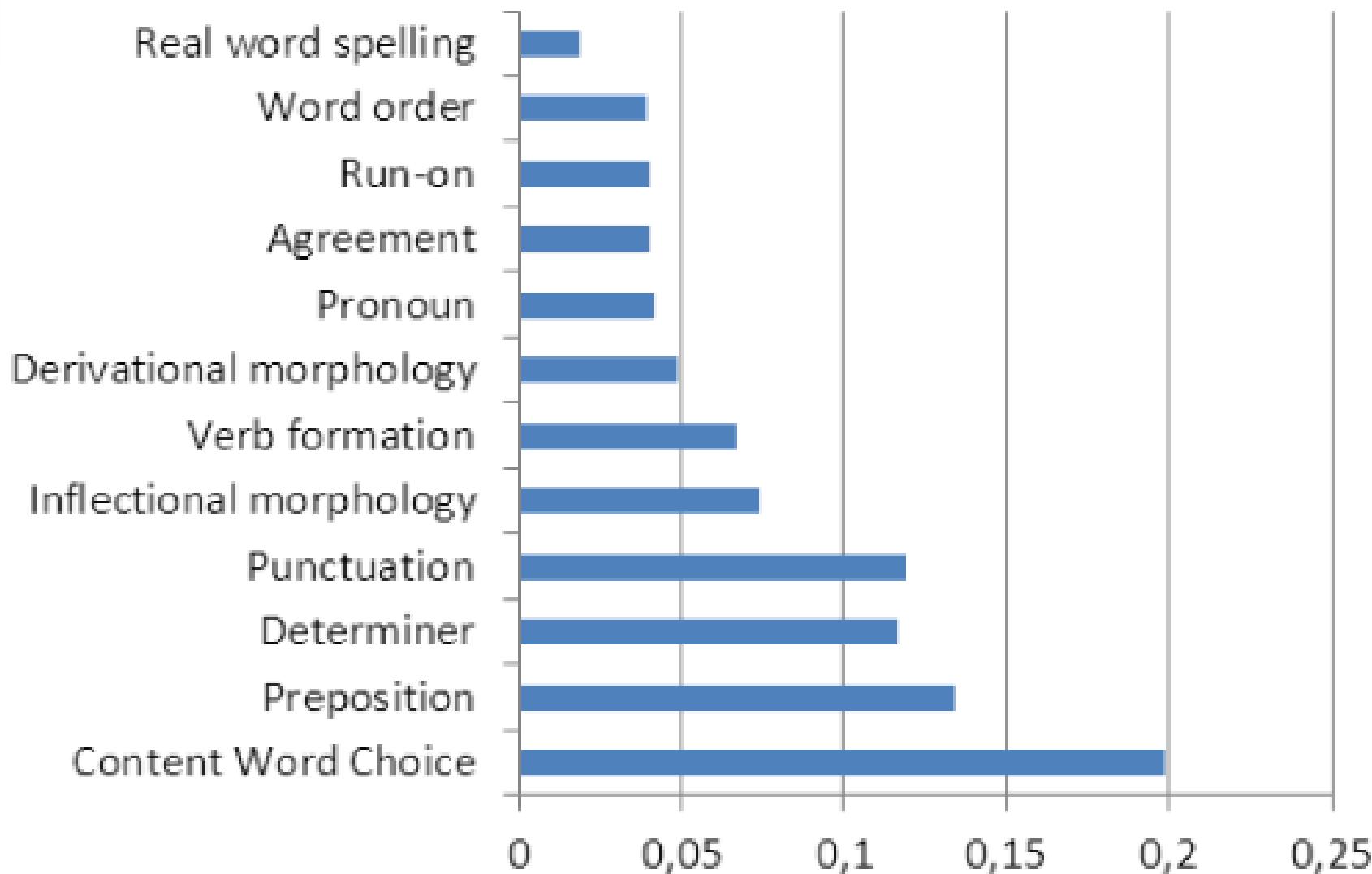
Focus on English only for practical reasons

- Over a billion people speak English as a second or foreign language worldwide
- By 2025, estimated that English language learners will make up 25% of the US public school population
- 725,000 international students at US universities
- 627 million people have taken the TOEFL
- English language has most resources

Learner errors



Learner errors



Prepositions Presence and Choice: 13%

- Prepositions are problematic because they perform so many complex roles
- Preposition choice in an adjunct is constrained by its object (“leave on Friday”, “leave at noon”)
- Prepositions are used to mark the arguments of a predicate (“fond of beer.”)
- Phrasal Verbs (“give in to their demands.”)
- “give in” -> “acquiesce”, “surrender”



Preposition choice

- Multiple prepositions can appear in the same context:

“When the plant is horizontal, the force of the gravity causes the sap to move __ the underside of the stem.”

Choices	Source
<ul style="list-style-type: none">• to• on• toward• onto	<ul style="list-style-type: none">• Writer• System• Annotator 1• Annotator 2

Determiner Presence and Choice: 12%

- English Article system: a, an, the
 - levels of countability from a car to *an equipment
 - Syntactic properties: have a knowledge vs a knowledge of English
 - Discourse factors – previous mention
 - Idioms: kick the/a bucket
 - World Knowledge
 - the moon (on earth)



Punctuation Conventions

- Apostrophe (1%):
 - Possessives
 - Contractions
- Comma (10%)
 - Missing after introductory clause
- Hyphenation (1%)
 - when used adjectively



Verbal Morphology and Tense: 14%

- Over-regularization of irregular verbs
 - The women *weared/wore long dresses.
- Ill-formed tense, participle, infinitive, modal & auxiliary
 - I look forward to *see/seeing you.
 - People would *said/say
 - It can *do/be harmful.
- Can be dependent on discourse
 - I will clean my room yesterday



Derivational Morphology: 5%

- Confusion of adjectival, nominal, verbal, adverbial forms
 - I have already made the *arranged/arrangements.
 - There was a wonderful women volleyball match between Chinese team and *Cuba/Cuban team.
 - I *admiration/admire my teacher.



Pronoun Error: 4%

- **Use of wrong case Wrong gender**
 - *Him/He went to the store.
- Wrong gender
 - I met Jane and he showed me where to go.
- Vague pronoun reference
 - I'll position the target, and when I nod my head, shoot at it.

Agreement Error: 4%

- These can be long distance Subject-verb agreement:
 - Three new texts which deal with this problem *has/have been written last year.
- Subject-verb agreement:
 - I *were/was in my house.
- Noun-number agreement
 - I am reading *these/this book.
 - Conversion always takes a lot of *efforts/effort.

Run-on Sentences: 4%

- Two independent clauses not connected by appropriate punctuation or conjunction:
 - They deliver documents to them they provide fast service.
 - It is nearly half past five, we cannot reach town before dark.



Word Order (4%)

- **Idiomatic Ordering of adjectives & nominal compounds**
 - tried and true vs true and tried
 - Ordering of adjectives & nominal compounds
 - A pop British band called “Spice Girl”.
 - English word order: subject verb object (SVO) Eat kids free (VSO)

Real Word Spelling Errors (2%)

- Homophones
 - there, their, they're
 - to, too, two
- Near Homophones
 - affect, effect
 - lose, loose



Content Word Choice: 20%

- Most common & least understood.
- Cover a wide range of errors & not fall into a pattern
- Collocation:
- strong / *powerful tea*strong / powerful computer
- Confusion of similar looking or sounding words:
 - Deliver the merchandise on a daily *base/basis



Influence of the Native Language

- L1 has no close equivalent construction
 - leading to difficulty in learning
 - Chinese and Russian have no equivalent of articles
- L1 has close equivalent construction
 - Easier to learn.
 - German article system similar to English
- Two languages closely related
 - transfer problems where they differ
- Unrelated languages
 - no transfer but will make more errors due to difficulty of complex English structures
 - Chinese/Russians need to learn the article rules

Goal of Grammatical Error Correction

- Grammatical error correction systems, like Microsoft Word, cover error types made by native speakers.
- They rarely identify article or preposition errors.
- Need systems that focus on those problems made by Language Learners: eg, articles, prepositions, verb formation, collocations, content word choice ...

Different Approaches

A: Rule-Driven:
No Context Needed

C: Parsing: Require syntactic
structure,
in sentence and beyond

E: Whole Sentence
Correction



B: Rule-Driven:
Local Context Needed

D: Machine Learning
methods

A: No Context Needed: Simple as a Regular Expression

- Regular expressions for many verb errors:

- Infinitive formation

$/to(RB)^* VB[DNGZ]/ \rightarrow /to(RB)^* \text{talk}/$

to talking → to talk

- Modal verb + have + past participle

$/MD \text{ of } VBD/ \rightarrow /MD \text{ have } VBD/$

would of liked → would have liked

- Word lists

- Over-regularized morphology: I eated/ate an omelet



Background: NLP Tools

- The chickens are crossing the road.

Token	The	chickens	are	crossing	the	road	.
Lemma	the	chicken	be	cross	the	road	.
POS1	DET	NOUN	VERB	VERB	DET	NOUN	PUNCT
POS2	DT	NNS	VBP	VBG	DT	NN	.
Parse	det	nsubj	aux	ROOT	det	dobj	punct

1. Rule-Based Methods

✍ Subject-verb agreement



The **chicken** crosses the road .

The **chicken** cross the road .

The **chickens** crosses the road .

The **chickens** cross the road .

1. Find subject of verb: *nsubj*

2. If NN VBZ **No error**

NN VBP **Error**

NNS VBZ **Error**

NNS VBP **No error**

3. Generate correction from
lemma + number

1. Rule-Based Methods

Complications

The **chicken cross** the road.

The **chicken are crossing** the road.

The **chicken and the duck is** crossing the road.

The **chicken looks both ways and then cross** the road.

The **chicken which hate** cars crosses the road.

Crossing the road **are** dangerous for a chicken.

The **police is** outside.

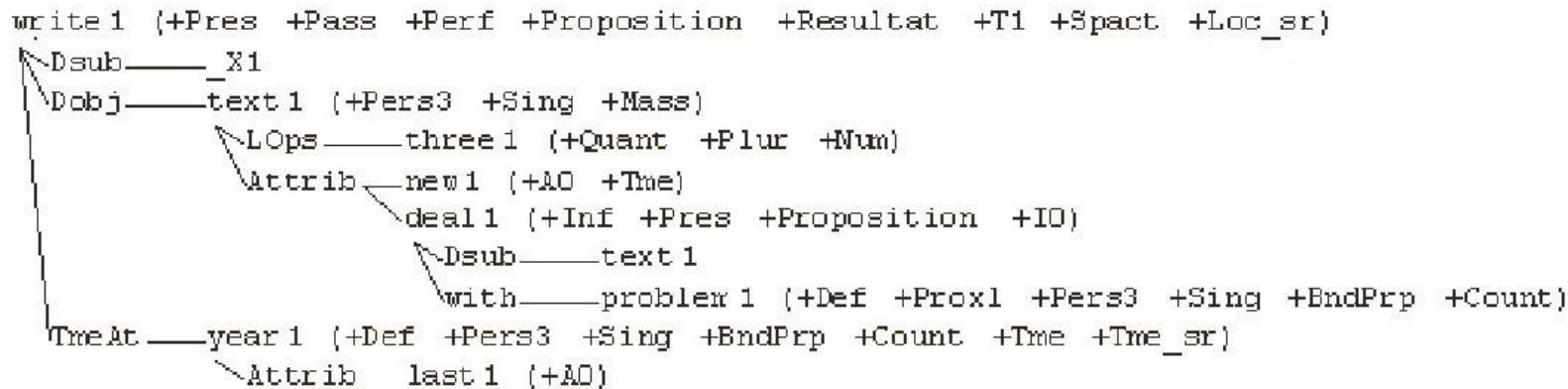
Five **pounds are** a lot of money.

C: Parsing/Logical Form

- 1980's: Before statistical parsers, modified parsers to recognize targeted errors
 - Allow parse trees that violate constraints – increment counter. Best solution has smallest index.
 - Add weights for specific violations
 - Mal-rules: Write rules to detect specific errors
 - Relax constraints on feature unification & use violations to produce feedback
- None allow for analysis of arbitrarily ungrammatical text

C: Microsoft Word

Complex hand-crafted phrase structure rules that read off of a logical form.



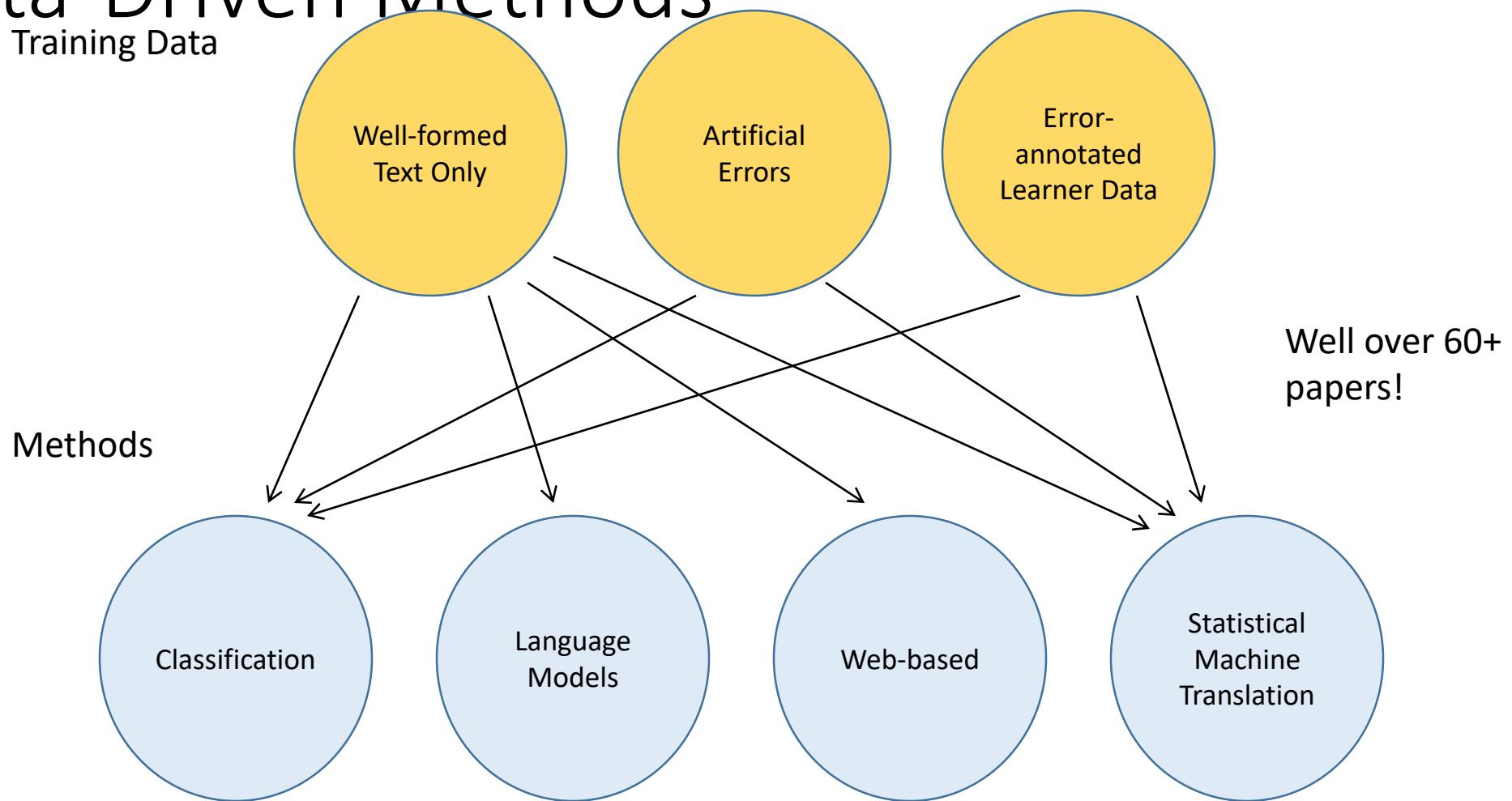
- Parse: Three new text which deal with the problem has been written.
- Pl quantifier and sg head noun. Suggest: text → texts
- Parse: Three new text which deal with the problem has been written.
- Detect subject-verb disagreement. Suggest: has → have
- Parse: no error detected

D: Error types that Require Data-Driven Methods

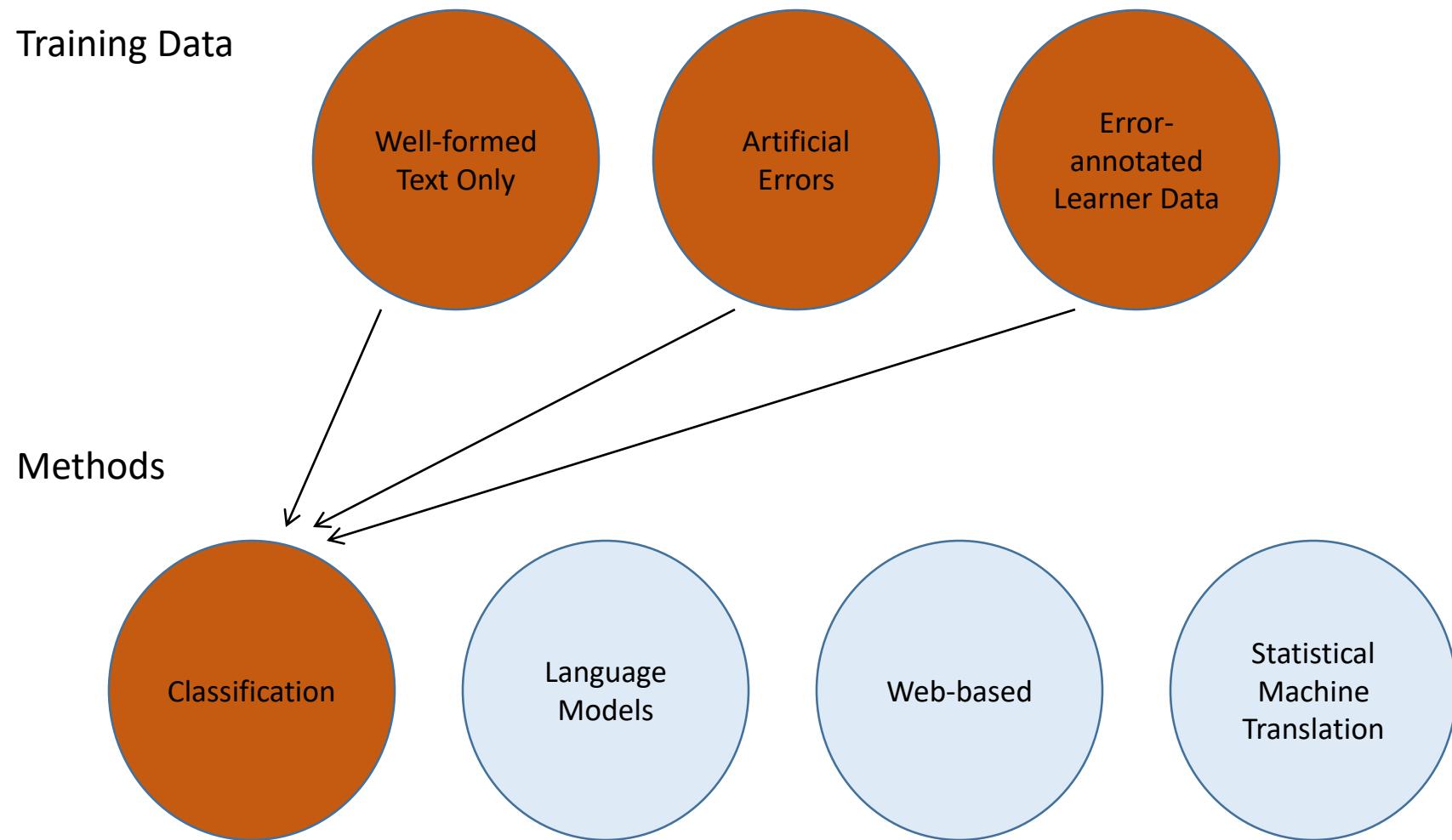
- Articles (*a, an, the*): presence and choice
- Prepositions (10 – 27): presence and choice
- Auxiliary verbs (*be, do, have*): presence and choice
 - A fire will break out and it can do/*be harm to people
 - A fire will break out and it can *do/be harmful to people.
- Gerund/Infinitive Confusion
 - On Saturday, I with my classmate went *eating/to eat.
 - Money is important in improving/*improve people's spirit.
- All verb errors – Lee & Seneff (2008), Rozovskaya et al (2014)

Data-Driven Methods

Training Data



Data-Driven Methods



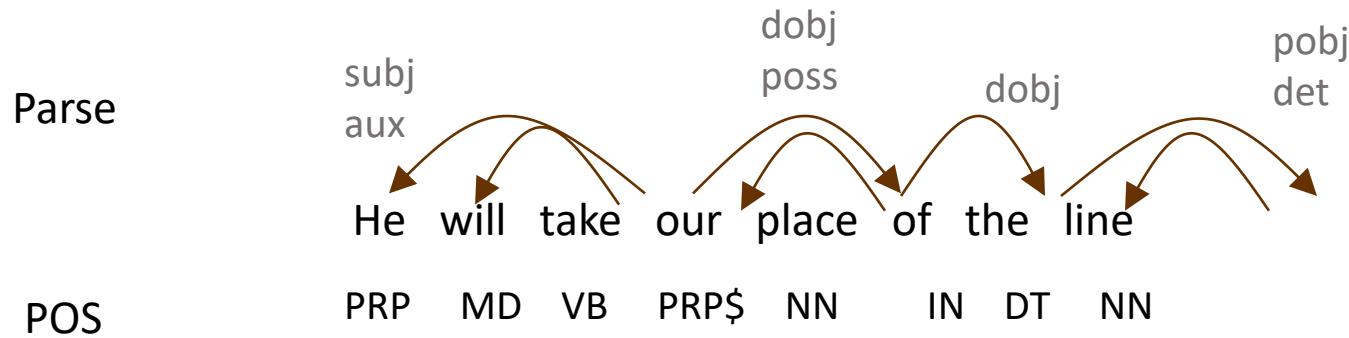
D: Data-Driven Methods

- *Supervised classification* requires:
 - Machine learning classifier (MaxEnt, SVM, Average Perceptron, etc.)
 - Data with labels for each training example

Label	Example
Correct	He will take our place in the line.
Error	He will take our place of the line.

Also need features!

Typical Features



WordNet
VerbNet
NER taggers
Semantic Role Labelers

Semantic

1-gram: place, the
2-gram: our-place, place-of, of-the, the-line
3-gram: our-place-of, place-of-the, of-the-line

Source

Writer's word(s) selection
L1 of writer
Genre of writing

N-grams

Types of Training Data

1. Training on examples of correct usage *only*
2. Training on examples of correct usage and *artificially generated errors*
3. Training on examples of correct usage and *real learner errors*

Choice of training data largely determined by availability of data

1. Training on Correct Usage

- Prior to 2010, very few error-annotated corpora to get enough examples of errors for ML
- Solution: train on examples of correct usage only
 - [Han et al., 2006; Tetreault and Chodorow, 2008; Gamon et al., 2008; Felice and Pulman, 2009]
- Advantages: plenty of well-formed text available
 - Google n-gram corpus to build language models
 - Large corpora such as news, Wikipedia, etc. to derive features from
- Challenges:
 - Best to match genre of learner writing, so need lots of well-formed student essays
 - Does not exploit any information of when or how errors tend to appear

2. Artificial Errors



- Training only on examples of correct usage has performance limitations
- Approximate learner writing by introducing artificial errors into a corpus of well-formed text
- Training instances
 - “Positive”: well-formed text
 - “Negative”: artificial errors
- Add a feature to capture transformation from erroneous choice to correct choice
- Challenge: determining the best way to approximate the errors

3. Error-Annotated Corpora

- Most common approach in shared tasks now that there are some labeled corpora available
- Use writer's word choice as a feature
- Some key works:
 - Han et al. (2010): showed that having a large corpus of annotated essays significantly outperformed positive-examples-only training on prepositions
 - Dahlmeier & Ng (2011): showed that Alternating Optimization Techniques worked well with error-annotated data for prepositions
 - Most CoNLL 2014 shared task systems

Classification

- ✍ They **were eaten** ice-cream when I **arrived**.
- ✍ Goal: Predict the correct form of a verb.

Verb Forms	Part-Of-Speech	Example
Infinitive	VB	eat
Past Tense	VBD	ate
Gerund or Present Participle	VBG	eating
Past Participle	VBN	eaten
Non-3 rd Person Singular Present	VBP	eat
3 rd Person Singular Present	VBZ	eats

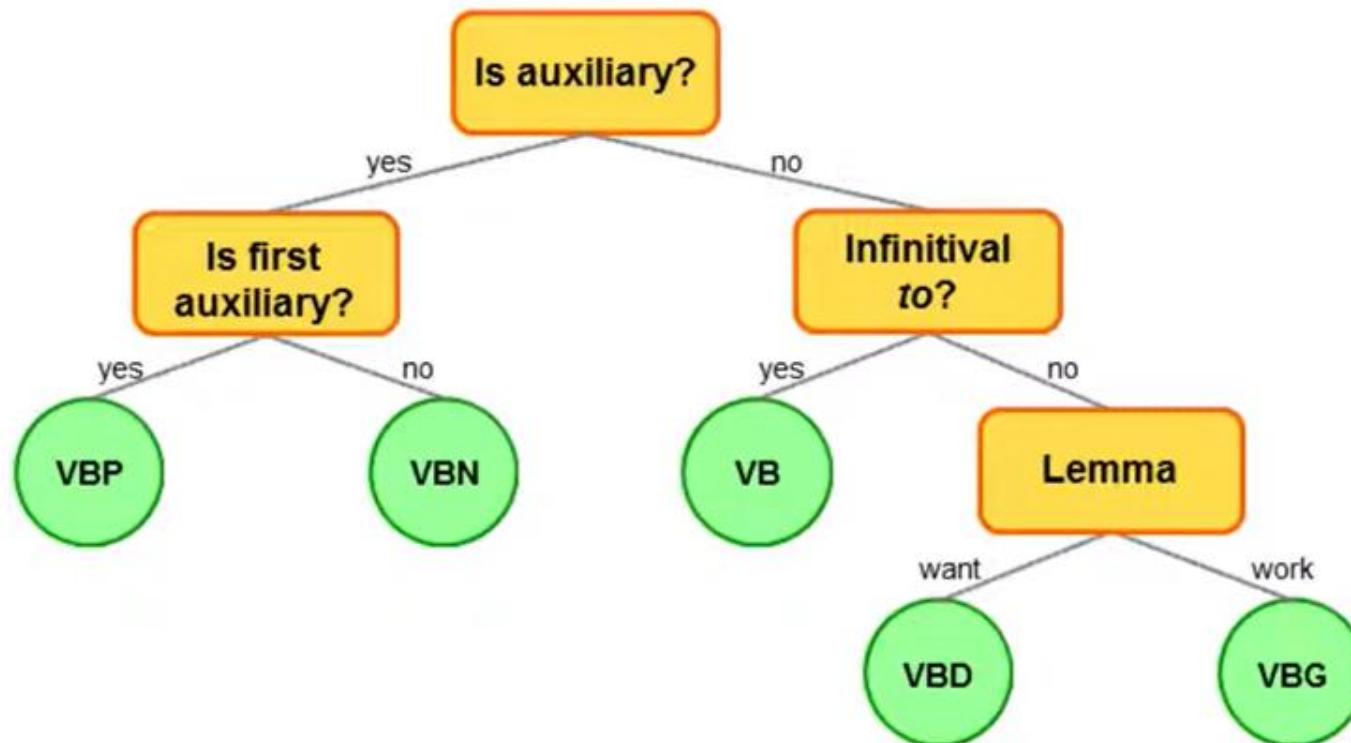
Classification - Features

- ✍ They **were eaten** ice-cream when I **arrived**.
- ✍ Instead of rules, define features!

Sample features	Values	Example: “were”
Is auxiliary?	Yes/No	Yes
Is first auxiliary?	Yes/No	Yes
Infinitival to?	Yes/No	No
Is ROOT verb?	Yes/No	No
Lemma	Lemma	be
N-gram	N words	They were eaten
Time expressions?	today, yesterday, ...	when ...
Target	VB, VBZ, VBP, ...	VBD

Classification - training

- ✍ Use machine learning to weight each feature.
 - Logistic Regression, Decision Trees, Support Vector Machines, ...



Classification

Advantages

- More flexible than rules.
- Incorporates human intuition.
- Only requires native data.

Disadvantages

- Feature engineering can be complicated.
- Better for smaller confusion sets.
- Only targets single error types.

Language Models

- ✍ Which is “best”?
 - Why did the chicken cross the **the**?
 - Why did the chicken cross the **line**?
 - Why did the chicken cross the **street**?
 - Why did the chicken cross the **road**?
- ✍ In a given context, some words more probable than others.
- ✍ Use this property to flag improbable sequences as errors.

Language Models

- ✍ N-gram models

1	Why	did	the	...
2	Why did	did the	the chicken	...
3	Why did the	did the chicken	the chicken cross	...
4	Why did the chicken	did the chicken cross	the chicken cross the	...

- ✍ Low N: High frequency, small context

- ✍ High N: Low frequency, large context

- ✍ Bigram (n=2) sentence probability

$$\rightarrow P(\text{Why}|\langle s \rangle) \times P(\text{did}|\text{Why}) \times P(\text{the}|\text{did}) \times P(\text{chicken}|\text{the}) \times \dots$$

Language Models

- ✍ Train a model on native text; e.g. Wikipedia
- ✍ Generate a confusion set
 - in, at, from, on, with, ...
- ✍ Score alternatives to find the best
 - I often work **in** home. |
 - I often work **at** home. ━━
 - I often work **from** home. ━━━
 - I often work **on** home. |
 - I often work **with** home. |
 - ...

Language Models

- Targeted Approach: can use LM scores over phrase or sentence for correction and detection

at	0.1
by	0.2
for	0.1
He will take our place in the line.	→ 0.3
from	0.0
to	0.1
with	0.1

- Similar to Web-based approach though one has more control of the data
- Nearly half of the HOO2012 systems used LMs

Language Models - Advantages

- ☞ Only require (lots of) native data
- ☞ Fairly easy to implement
- ☞ Can detect all error types
- ☞ Collocation errors
 - I have a **strong/powerful** computer.
 - This **effected/affected** me deeply.
- ☞ Versatile

Language Models - Disadvantages

- ☞ Probability is not grammaticality.
 - Student: I is ...
 - Teacher: No, you must always say “I am”.
 - Student: Ok. I am the ninth letter of the alphabet.

- I am at home. 
- I was at home. 

- ☞ Rare/unseen words: paraklausithyron, covfefe
- ☞ Long range dependencies
 - The house on the corner with the blue windows is new.

Methods: Web-Based Methods

- Language learners will typically look at counts returned by search engine to figure out best word to use
- What happens when we use this simple methodology?
 - Select “target word” and search for alternatives
 - Select alternative with top web count

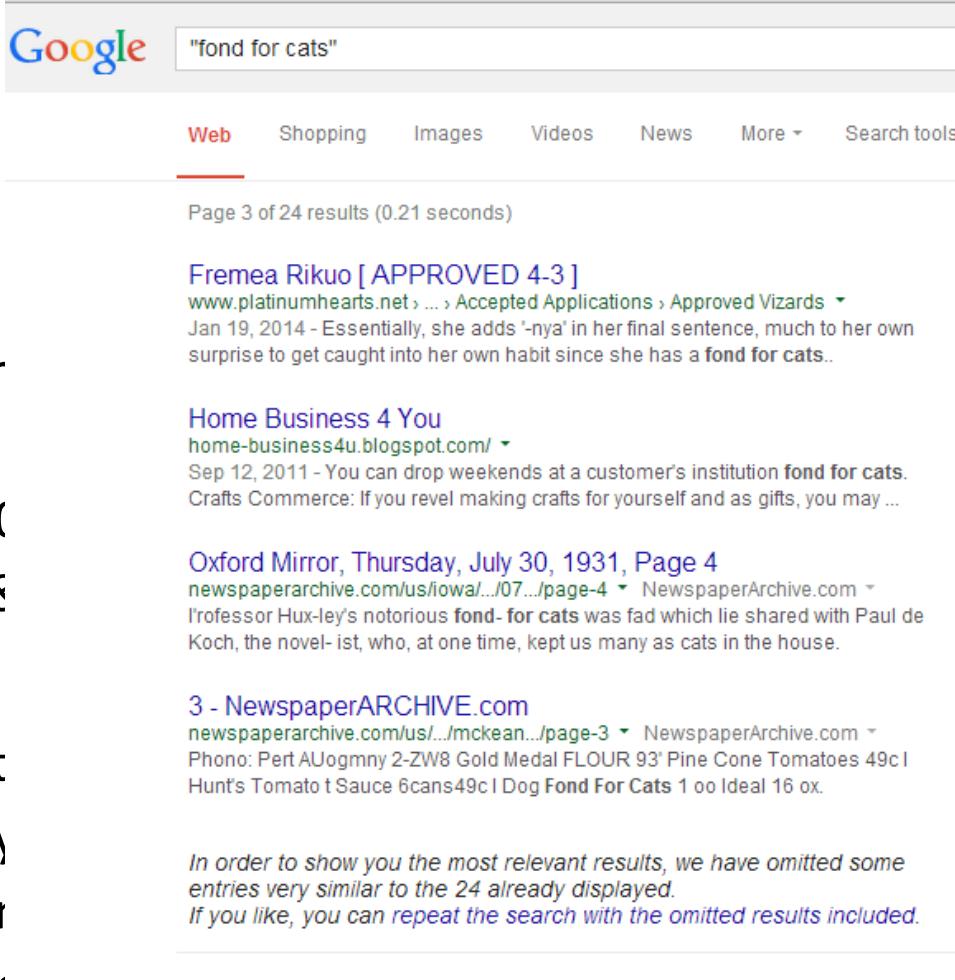
Web-Based Methods

Phrase	Google Count	Bing Count
“fond of cats”	638,000	42,800
“fond for cats”	178	2
“fond by cats”	0	0
“fond to cats”	269	5
“fond with cats”	13,300	10



Method

- Prior work shows approaches
 - Yi et al. (2011)
 - Tetreault & Smith (2012)
- Issues:
 1. No POS tagging
 2. Search syntax
 3. Constraint
 4. Search counts are for pages not instances
 5. Search engines behave differently



A screenshot of a Google search results page for the query "fond for cats". The search bar at the top contains the query. Below it, the "Web" tab is selected. The results show four entries:

- Fremea Rikuo [APPROVED 4-3]**
www.platinumhearts.net › ... › Accepted Applications › Approved Wizards ›
Jan 19, 2014 - Essentially, she adds '-nya' in her final sentence, much to her own surprise to get caught into her own habit since she has a fond for cats..
- Home Business 4 You**
home-business4u.blogspot.com/ ›
Sep 12, 2011 - You can drop weekends at a customer's institution fond for cats. Crafts Commerce: If you revel making crafts for yourself and as gifts, you may ...
- Oxford Mirror, Thursday, July 30, 1931, Page 4**
newspaperarchive.com/us/iowa/.../07.../page-4 › NewspaperArchive.com ›
Professor Huxley's notorious fond-for cats was fad which he shared with Paul de Koch, the novelist, who, at one time, kept us many as cats in the house.
- 3 - NewspaperARCHIVE.com**
newspaperarchive.com/us/.../mckean.../page-3 › NewspaperArchive.com ›
Phono: Pert AUogmny 2-ZW8 Gold Medal FLOUR 93' Pine Cone Tomatoes 49c | Hunt's Tomato t Sauce 6cans49c | Dog Fond For Cats 1 oo Ideal 16 ox.

In the bottom right corner of the search results, there is a note: *In order to show you the most relevant results, we have omitted some entries very similar to the 24 already displayed. If you like, you can repeat the search with the omitted results included.*

ods

ot over classification
tion of queries
nine L1 specific errors

Statistical Machine Translation

✍ Greek to English

Άφησα το βιβλίο μου στο σπίτι .
I left my book at home .

✍ “Bad” English to “Good” English

I leaved my book in the home .
I left my book at home .

Statistical Machine Translation

- ✍ Requires parallel data

Name	Type	# Words	Error Types?	Availability
CLC	Essays	~29 million	Yes (80)	Private
Lang 8	Web Forum	~12 million	No	Public
NUCLE	Essays	~1.2 million	Yes (28)	Public

- ✍ Issues

- >10x smaller than translation corpora
- Annotation quality
- Proficiency levels
- Genre

Error-Annotated Corpora

NUCLE

- **National University of Singapore Corpus of English**
- 1,450 essays by Singapore college students
- Used in CoNLL shared tasks
- Publically available

FCE

- 1,244 essays from **First Certificate in English** exam (CLC subset)
- Used in HOO 2012 task
- Includes score, error annotation and demographics
- Publically available

HOO2011

- Hand corrected papers from **ACL Anthology**
- 38 conference papers
- Publically available

CLEC

- **Chinese Learners of English Corpus**
- 1M words
- Five proficiency levels
- Inexpensive

Other Learner Corpora

TOEFL11

- **ETS Corpus of Non-Native English**
- 12,100 essays (1,100 essays each for 11 different L1s)
- Includes proficiency information
- Available through LDC

ICLE

- **International Corpus of Learner English**
- 3.7 M from over 16 different L1s
- Partially error-annotated
- Needs to be licensed.

Lang-8

- **Language Learner Social Community Website**
- Nearly 200,000 Learner journal entries with community corrections
- Need a script to extract data (Mizumoto et al., 2011)



mestre

Jul 26, 2014 12:47

- ✍ I'm going to an afternoon mini-concert that's going to take place at the small stage in the shopping mall.
- ✓ I'm going to an afternoon mini-concert that's going to take place at a small stage in the shopping mall.

"the" may be correct here, but since we're not familiar with the shopping mall in question, "a" feels more natural.

1 people think this correction is good.

- ✍ My daughter in law is going to play the sax in it.
- ✓ My daughter in law will be playing the sax.

More of a suggestion than anything.

1 people think this correction is good.

- ✍ She's cool, active, and she's well into the last trimester of pregnancy, and proud.
- ✓ She's cool, active, and she's well into the last trimester of pregnancy, and proud.

This sentence sounds rather unnatural with the 'and' both here and at the end portion, so I removed this bit. Adding "of it" after "proud" will make that portion stronger and more specific, but may not be your intention.

1 people think this correction is good.

2
Entries Written

2
Entries Written

2
Entries Written

Latest entries

● A New Trend of University English Entrance Examination (9)

● A Mini-Concert (18)

[Read more](#)

Latest comments

● 07:03 いしぐろ Ishiguro○_○

● 05:22 Kalyak

● 04:50 tony

● 03:40 student113

● 00:38 建

Statistical Machine Translation

- ☛ Input: Parallel sentence pairs
- ☛ Phrase-based SMT learns the mapping between phrases.
 - [το βιβλίο μου → my book]
 - [discuss about → discuss]
 - [informations → information]
 - [as result → as a result]
 - [the → the]
- ☛ The mappings are stored in a *phrase table*.
- ☛ Translate using the phrase table and a language model.

Statistical Machine Translation

Goal: Find the best translation

Let 's discuss about this informations .

Let 's discuss about this informations .

Lets talk over the information ?

Let 's discuss the information !

talk about this information

these informations

Let 's discuss this information .

Two Classes of GEC / SMT

GEC=Grammer error check

1. Noisy Channel Model

- View error correction as the process of *translating* from learner English to fluent English

2. Round Trip Machine Translation

- View SMT as a “black box” and use MT engine to generate possible corrections

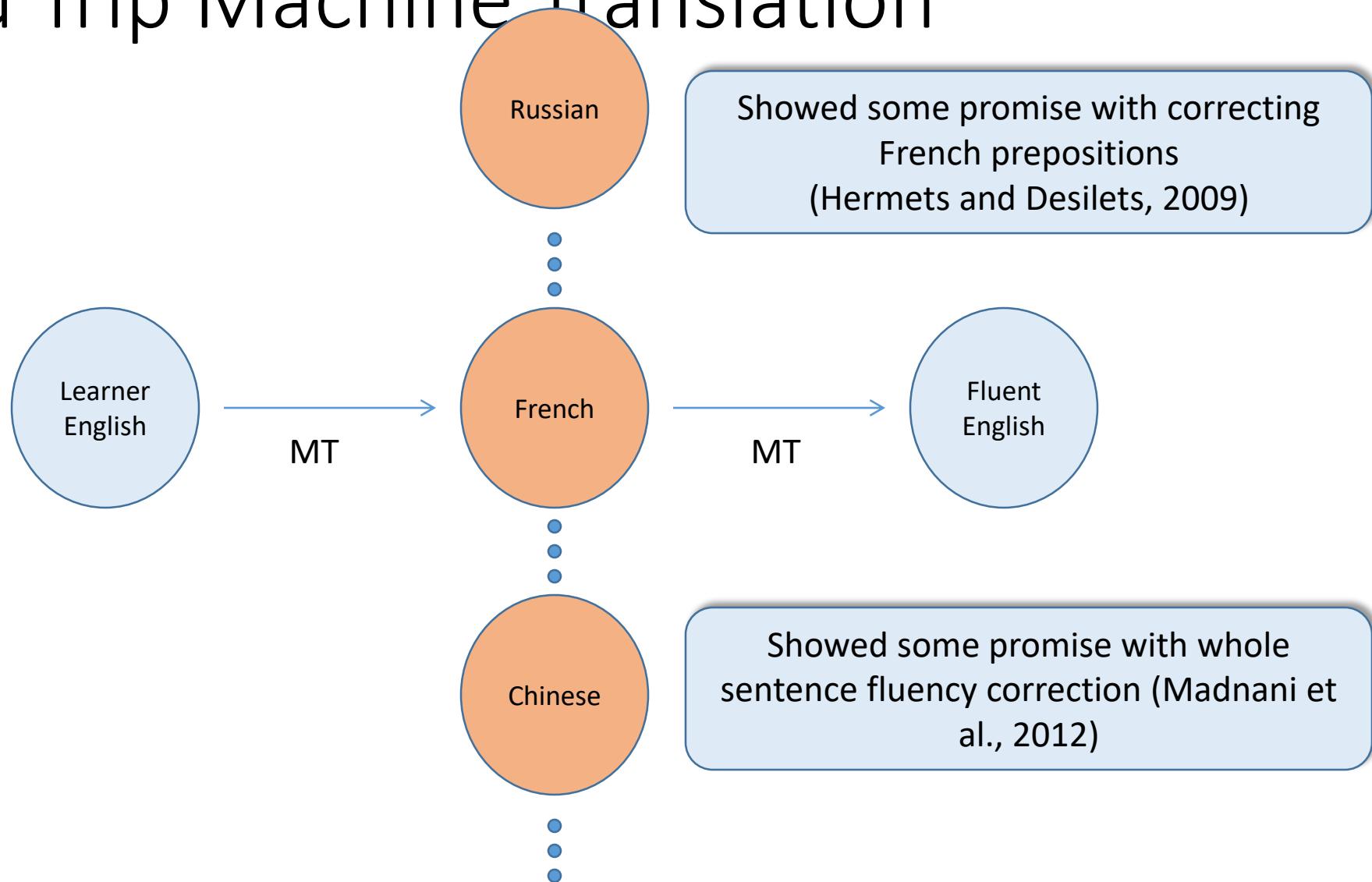
Noisy Channel Model

- Re-train MT system with examples of error phrases (or sentences) and their corrections
- Dependent on having enough error-annotated data
- Some examples:
 - Brocket et al. (2006): use artificial errors to train SMT to correct mass noun errors
 - Park & Levy (2011): use technique with FSTs

Round Trip Machine Translation

- Use pre-existing MT system to translate a sentence into another language and translate back into English
 - Thus does not use learner data
- Preliminary pilot studies with this method show some potential

Round Trip Machine Translation



Statistical Machine Translation



Advantages

- Can detect and correct all error types simultaneously
- Handles interacting errors
- No feature engineering or expert knowledge
- State-of-the-art



Disadvantages

- Long training time: several days
- Require lots of parallel training data
- Out-Of-Vocabulary words (OOV)
- Hard to customise

Traditional NLP Evaluation Metrics

- Precision, Recall and F-score are all used to evaluate shared tasks
- However they can be problematic for GEC evaluation and should be interpreted with caution

Traditional NLP Evaluation Metrics

$$Precision = \frac{TP_S}{TP_S + FP_S}$$

$$Recall = \frac{TP_S}{TP_S + FN_S}$$

$$F\text{-}score = \frac{Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP_S + TN_S}{TP_S + TN_S + FP_S + FN_S}$$

Terminology

- True Positive (TP) “hit”
 - Flags *I am going for walk this afternoon.*
- False Positive (FP)
 - Flags *I am going for a walk this afternoon.*
- True Negative (TN)
 - Not flag *I am going for a walk this afternoon.*
- False Negative (FN) “miss”
 - Not flag *I am going for walk this afternoon.*



Shared Tasks/Competitions

- Important for a field to progress
 - Helping Our Own (HOO): 2011 & 2012
 - Conference on Computational Natural Language Learning (CoNLL): 2013 & 2014
- Shared train and evaluation data sets
- Shared evaluation metrics

Shared Task	Errors	Corpus	# of Teams
HOO 2011	All	ACL Papers	6
HOO 2012	Preps & Dets	FCE / (CLC)	14
CoNLL 2013	Preps, Dets, Nouns, Verbs	NUCLE	17
CoNLL 2014	All	NUCLE	12

Shared Task Evaluation Metrics

- HOO 2011: Three evaluations
 - Detection: Identify error
 - Recognition: Identify an error's type and span
 - Correction: Provide at least one accurate rewrite
 - Precision, Recall & F-score calculated for each
- HOO 2012: Same as HOO 2011 plus
 - Participating teams could request changes in the annotation – adjudicated by organizers. Increased F-scores by almost 10%
- CoNLL: Same as HOO but different mapping

HOO Mapping

- HOO
 - Detection:
 - Any overlap with gold edit=TP
 - Output not overlap a gold edit=FP
 - No overlap with gold edit=FN
 - Recognition: edits must be exact
 - Correction: edits and labels must be exact

Shared Tasks: Lessons Learned

- Performance
 - Despite 4 tasks, performance low: 20 to 40 F-score
- Annotation Quality:
 - Inconsistent
 - Systems penalized for valid corrections not annotated
 - Last 3 shared tasks allowed revisions to annotations by participants
 - The revisions increased F-score by almost 10%
- Need to deal with multiple interacting errors.

Conclusion

- ❖ We looked at:

- Rule-based Methods
- Language Models
- Classifiers
- Machine Translation

- ❖ State of the art

- Detection: $P \approx 0.70$, $R \approx 0.30$
- Correction: $P \approx 0.60$, $R \approx 0.25$

Grammar Checking Approaches

1. Rule-Based Approaches:

- These methods employ predefined grammar rules and patterns to identify and correct errors in text.
- Grammar rules are formulated based on linguistic principles and syntax.
- For instance, identifying subject-verb agreement errors, incorrect verb tense usage, punctuation errors, and sentence fragments.

2. Statistical Approaches:

- Statistical methods leverage machine learning algorithms to identify grammatical errors by analyzing large amounts of text data.
- These models learn patterns and probabilities of correct language usage based on annotated corpora.
- They can detect errors based on deviations from statistically common linguistic structures.

3. Hybrid Approaches:

- Combining rule-based and statistical methods can enhance the accuracy of grammar checking.
- Hybrid systems utilize both rule-based grammar rules and statistical models to detect and correct errors, leveraging the strengths of each approach.

Deep Learning Approaches

Deep Learning Approaches:

- Neural network-based models, such as recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformers, have been employed for grammar checking.
- These models can learn complex linguistic patterns and relationships, aiding in detecting errors and suggesting corrections.

Error-Correcting Models:

- Some NLP models are specifically trained to detect and correct errors.
- For instance, sequence-to-sequence models are trained on pairs of correct and incorrect sentences, learning to generate corrections for the given erroneous sentences.

Language Models for Contextual Understanding:

- Recent advancements in large language models, like BERT, GPT, and their variants, have improved grammar checking by incorporating contextual understanding.
- These models can comprehend the context of sentences and suggest corrections based on the surrounding words.

Grammar Checking Approaches

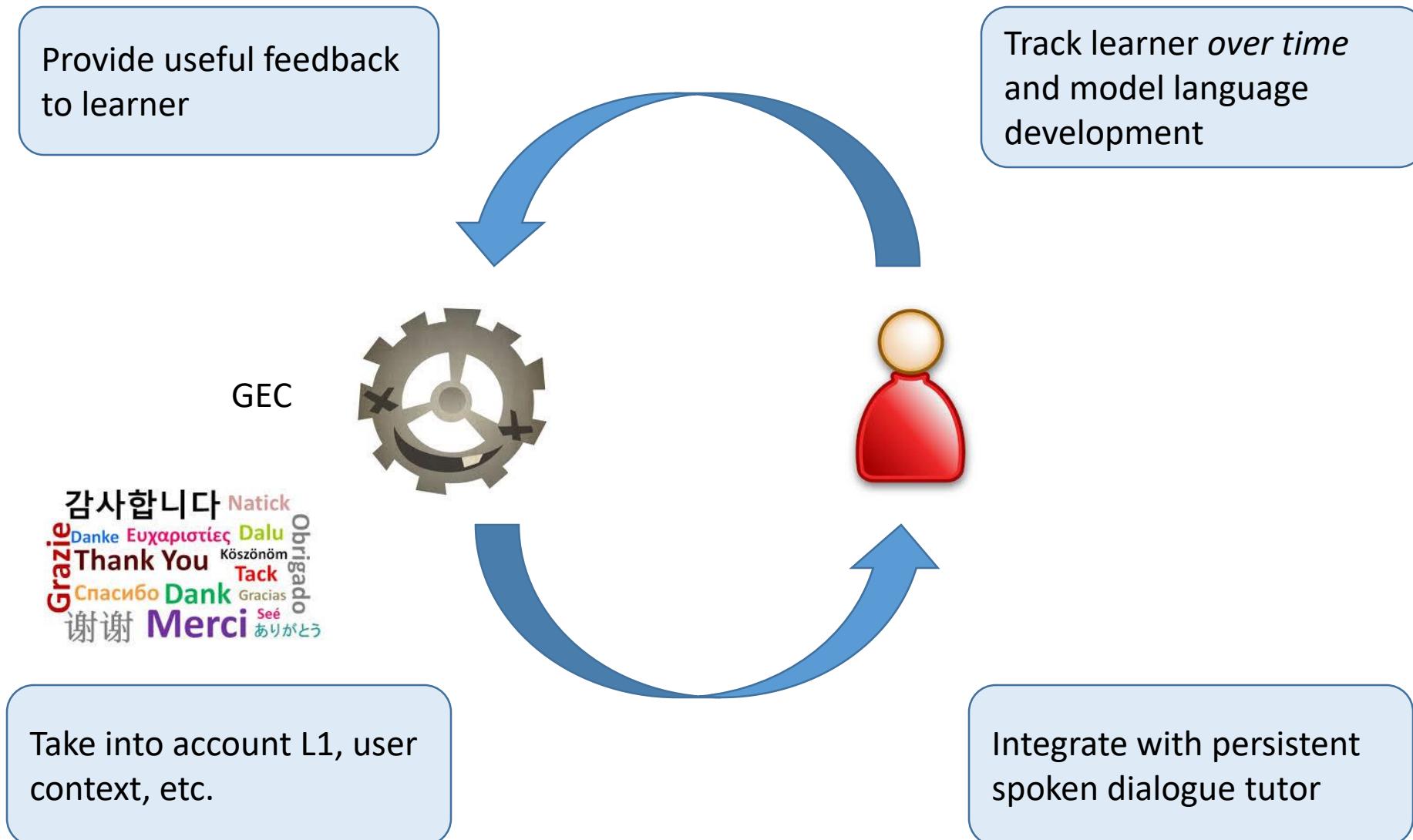
User Feedback and Machine Learning Iteration:

- Some grammar checkers improve over time by incorporating user feedback.
- Machine learning algorithms learn from user corrections to enhance their accuracy in identifying and correcting errors.

Fine-tuning and Transfer Learning:

- Pre-trained language models can be fine-tuned on specific grammar checking tasks by training them on annotated datasets.
- Transfer learning allows models to adapt their knowledge from general language understanding to the specific task of grammar checking.
- Transfer learning, used in machine learning, is the reuse of a pre-trained model on a new problem. In transfer learning, a machine exploits the knowledge gained from a previous task to improve generalization about another.

What is the future of GEC?



What is the future of GEC?

- A system which can automatically transform one noisy sentence to a fluent sentence...without a change in meaning

Having discuss all this I must say that I must rather prefer to be a leader than just a member.



GEC

After discussing all this I must say that I'd prefer to be a leader than a follower.

What is the future of GEC?

- System need not simply be a text to text transformation, could also take into account:
 - Other sentences in document
 - Context of document (writer's intention)
 - Register
 - Who the document is for
 - Prior sentences writer has produced (personalization)

Applications of GEC

- Automated Essay Scoring
- Native Language Identification
- MT Quality Estimation
- Noisy data processing
 - Social Media / normalization
 - MT post-processing
 - Assistive Tech: GEC of automatic closed captions

Summary

- Provided a history of GEC
- Described popular methodologies for correcting language learner errors
- Grammatical Error Correction one of the oldest fields and applications of NLP
- Still much work to be done as performance is still low!

References

- https://nlp.johnsnowlabs.com/analyze_spelling_grammar
 - <https://direct.mit.edu/coli/article/49/3/643/115846/Grammatical-Error-Correction-A-Survey-of-the-State>
 - <https://blogs.nvidia.com/blog/microsoft-triton-ai-grammar-word/>
 - <https://paperswithcode.com/datasets?q=&v=lst&o=newest&task=grammatical-error-correction&mod=texts&page=1>
 - <https://onlinelibrary.wiley.com/doi/10.1002/spe.2110>
 - <https://www.youtube.com/watch?v=3rVn14m8zaM>
 - <https://www.youtube.com/watch?v=pYV8OydsnQA>
 - <https://www.youtube.com/watch?v=9ZkM-jn-EWA>
 - <https://www.hindawi.com/journals/sp/2021/4213791/>
 - <https://www.youtube.com/watch?v=3rVn14m8zaM>
 - <https://sapling.ai/grammar-check>
-



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in





Session 4 : Question Answering

Date – 17th Dec 2023

Time – 1.40 pm to 2.40 pm

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Dan Jurafsky and many others who made their course materials freely available online.

Session Content

- Types of Questions
 - IR-based Factoid Question Answering
 - Knowledge-based Question Answering
 - Case Study: IBM's Watson
 - Deep Learning Algorithms
 - Hybrid Approach
 - Evaluation Metrics
-

Question Answering



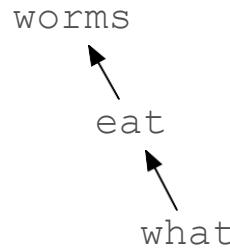
One of the oldest NLP tasks (punched card systems in 1961)

Question:

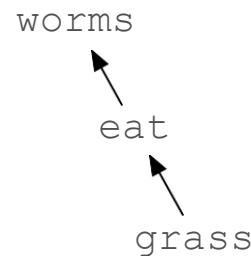
Potential Answers:

Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196–204

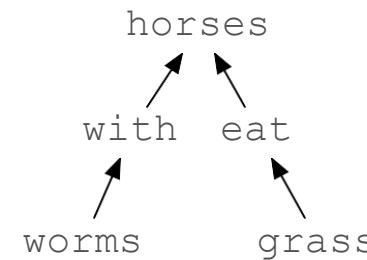
What do worms eat?



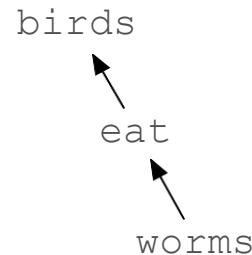
Worms eat grass



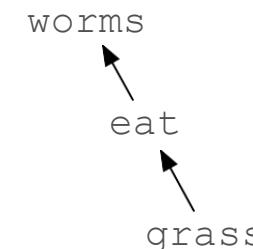
Horses with worms eat grass



Birds eat worms



Grass is eaten by worms



Question Answering: IBM's Watson



- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
“AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA”
INSPIRED THIS AUTHOR’S
MOST FAMOUS NOVEL



Bram Stoker

Apple's Siri



 **WolframAlpha**TM computational... knowledge engine

how many calories are in two slices of banana cream pie? ≡

≡ Examples Random

Assuming any type of pie, banana cream | Use pie, banana cream, prepared from recipe or pie, banana cream, no-bake type, prepared from mix instead

Input interpretation:

pie	amount	2 slices	total calories
type		banana cream	

Average result: Show details

702 Cal (dietary Calories)

Types of Questions in Modern Systems



- Factoid questions
 - *Who wrote “The Universal Declaration of Human Rights”?*
 - *How many calories are there in two slices of apple pie?*
 - *What is the average age of the onset of autism?*
 - *Where is Apple Computer based?*
- Complex (narrative) questions:
 - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
 - *What do scholars think about Jefferson’s position on dealing with pirates?*

Commercial systems: mainly factoid questions

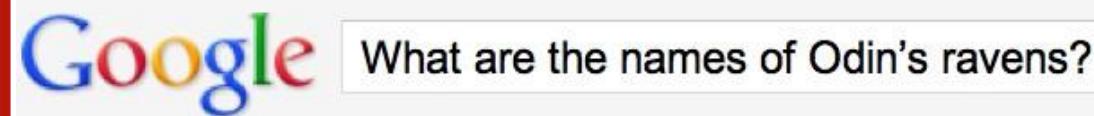
Where is the Louvre Museum located?	In Paris, France
What's the abbreviation for limited partnership?	L.P.
What are the names of Odin's ravens?	Huginn and Muninn
What currency is used in China?	The yuan
What kind of nuts are used in marzipan?	almonds
What instrument does Max Roach play?	drums
What is the telephone number for Stanford University?	650–723–2300

Paradigms for QA



- IR-based approaches
 - TREC; IBM Watson; Google
- Knowledge-based and Hybrid approaches
 - IBM Watson; Apple Siri; Wolfram Alpha; True Knowledge Evi

Many questions can already be answered by web search



Search About 214,000 results (0.38 seconds)

Everything
Images
Maps
...

[Huginn and Muninn - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Huginn_and_Muninn
The **names** of the **ravens** are sometimes modernly anglicized as Hugin and Munin. In the Poetic Edda, a disguised **Odin** expresses that he fears that they may ...
[Attestations](#) - Archaeological record - Theories - See also

IR-based Question Answering



Where is the Louvre Museum located?

Search

About 904,000 results (0.30 seconds)

Everything

Best guess for Louvre Museum Location is Paris, France

Mentioned on at least 7 websites including [wikipedia.org](#), [answers.com](#) and [east-buc.k12.ia.us](#) - Show sources - Feedback

Images

[**Musée du Louvre - Wikipedia, the free encyclopedia**](#)

en.wikipedia.org/wiki/Musée_du_Louvre

Maps

Musée du Louvre is located in Paris. Location within Paris. Established, 1793. Location,

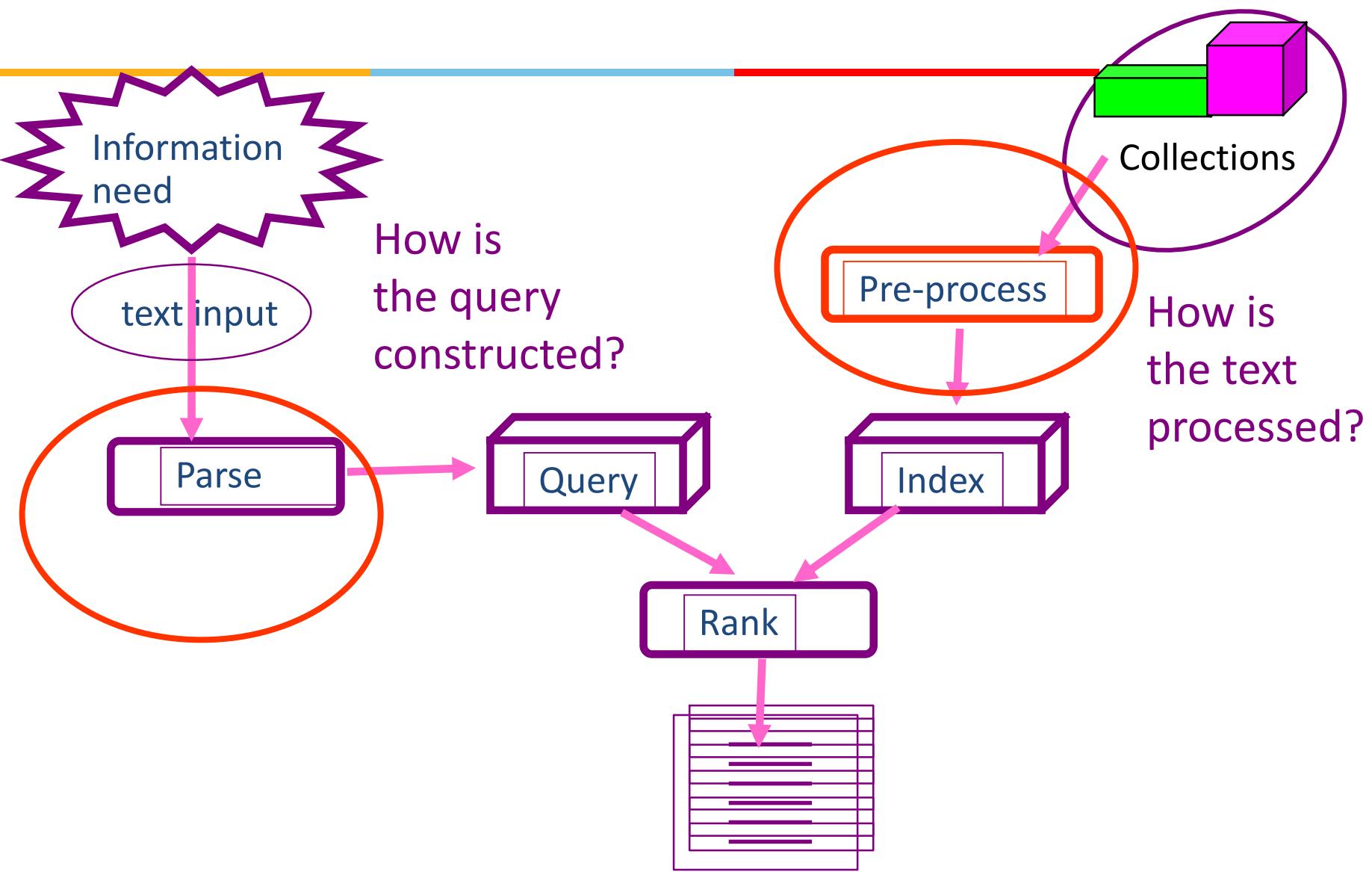
Palais Royal, Musée du Louvre, 75001 Paris, France. Type, Art museum ...

Videos

[Louvre Palace - List of works in the Louvre - Category:Musée du Louvre](#)

News

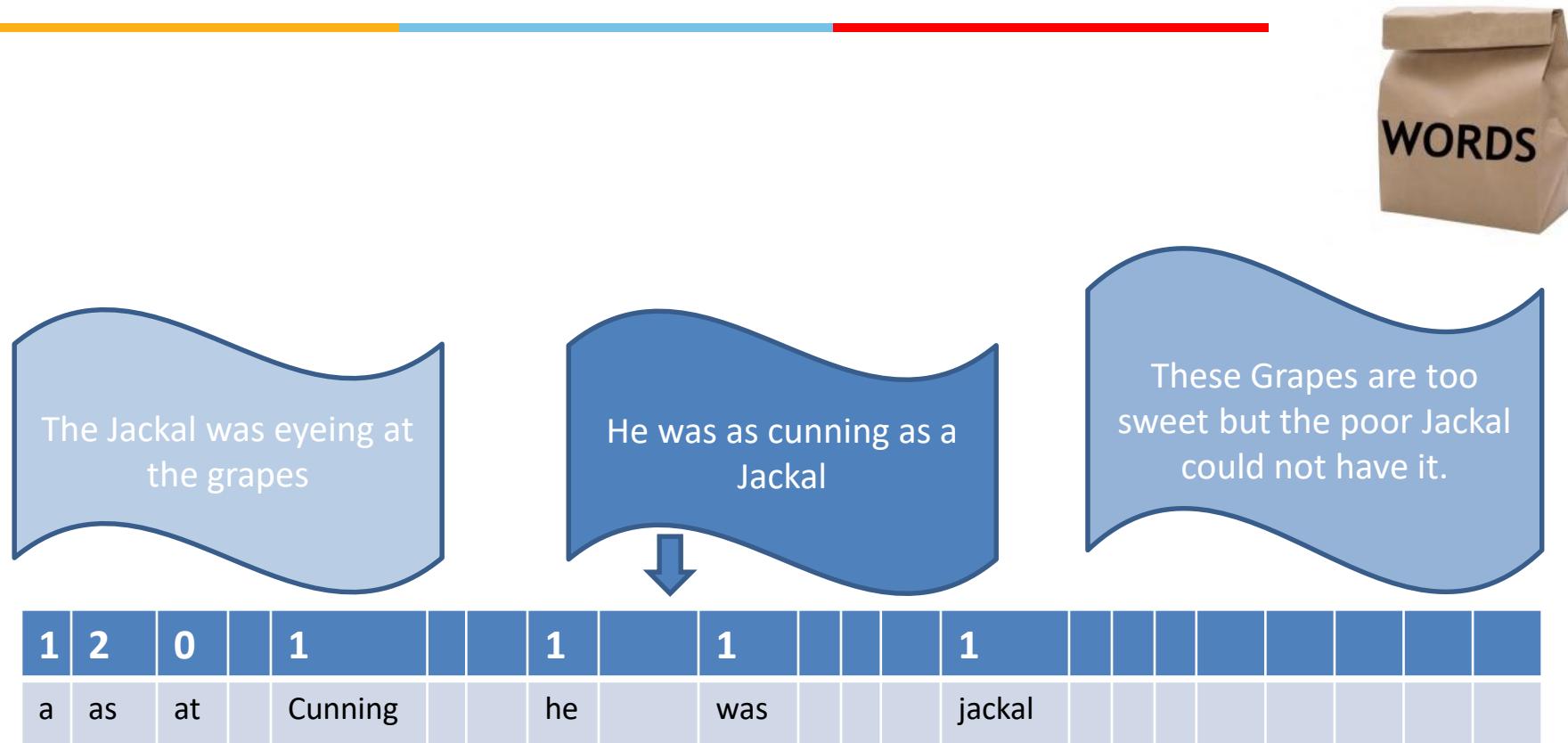
Information Retrieval Process



Bag of Words representation

- A very popular and basic representation of documents is the bag of words model.
- Each document is represented by a **bag (= multiset)** of **terms** from a predefined vocabulary.

Bag of Words representation



Term Incidence Matrix

	T1	T2	T3	T4	T5	T6
D1	1	0	0	1	1	0
D2	0	1	0	1	1	0
D3	1	0	1	0	1	1
D4	1	0	1	0	1	1

	T1	T2	T3	T4	T5	T6
D1	6	0	0	2	1	0
D2	0	8	0	5	3	0
D3	2	0	6	0	5	2
D4	5	0	2	0	6	7

Term Frequency

Frequency of occurrence for the term in each document is included in the vector

<i>docs</i>	<i>t1</i>	<i>t2</i>	<i>t3</i>
D1	2	0	3
D2	1	0	0
D3	0	4	7
D4	3	0	0
D5	1	6	3
D6	3	5	0
D7	0	8	0
D8	0	10	0
D9	0	0	1
D10	0	3	5
D11	4	0	1

Vector Space Model

Documents are represented as vectors in term space

- Terms are usually *stems*
- Documents represented by binary or weighted vectors of terms

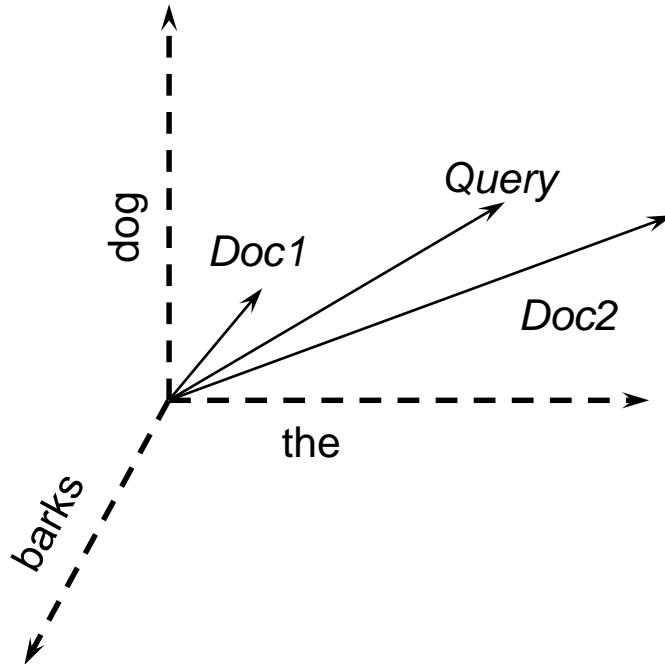
Queries represented the same as documents

Query and Document weights are based on length and direction of their vector

A vector distance measure between the query and documents is used to rank retrieved documents

Vector Coefficients

- How to represent the documents and queries?



Doc1: the dog barks <1 1 1>
Doc2: the dog dog barks barks
barks <1 2 3>

Query: the dog dog barks barks
<1 2 2>

Inverse document frequency

df_t is the document frequency of t : the number of documents that contain t

- df is a measure of the informativeness of t

We define the idf (inverse document frequency) of t by

$$\text{idf}_t = \log N/\text{df}_t$$

We use $\log N/\text{df}_t$ instead of N/df_t to “dampen” the effect of idf

Inverse Document Frequency

IDF provides high values for rare words and low values for common words

$$\log\left(\frac{10000}{10000}\right) = 0$$

For a collection of 10000 documents (N = 10000)

$$\log\left(\frac{10000}{5000}\right) = 0.301$$

$$\log\left(\frac{10000}{20}\right) = 2.698$$

$$\log\left(\frac{10000}{1}\right) = 4$$

Assigning Weights

tf*idf measure:

- Term frequency (tf)
- Inverse document frequency (idf)

Goal:

Assign a tf*idf weight to each term in each document

Simple tf*idf

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

T_k = term k in document D_i

tf_{ik} = frequency of term T_k in document D_i

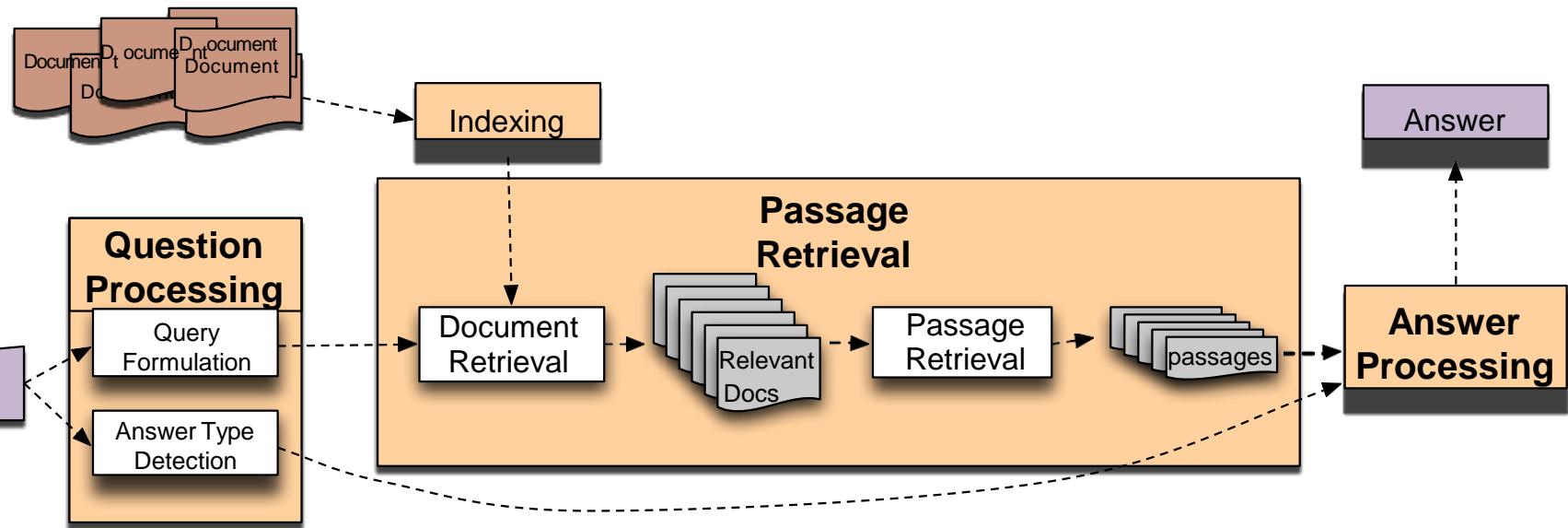
idf_k = inverse document frequency of term T_k in C

N = total number of documents in the collection C

n_k = the number of documents in C that contain T_k

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

IR-based Factoid QA



IR-based Factoid QA

- **QUESTION PROCESSING**
 - Detect question type, answer type, focus, relations
 - Formulate queries to send to a search engine
- **PASSAGE RETRIEVAL**
 - Retrieve ranked documents
 - Break into suitable passages and rerank
- **ANSWER PROCESSING**
 - Extract candidate answers
 - Rank candidates
 - using evidence from the text and external sources

Knowledge--based approaches (Siri)



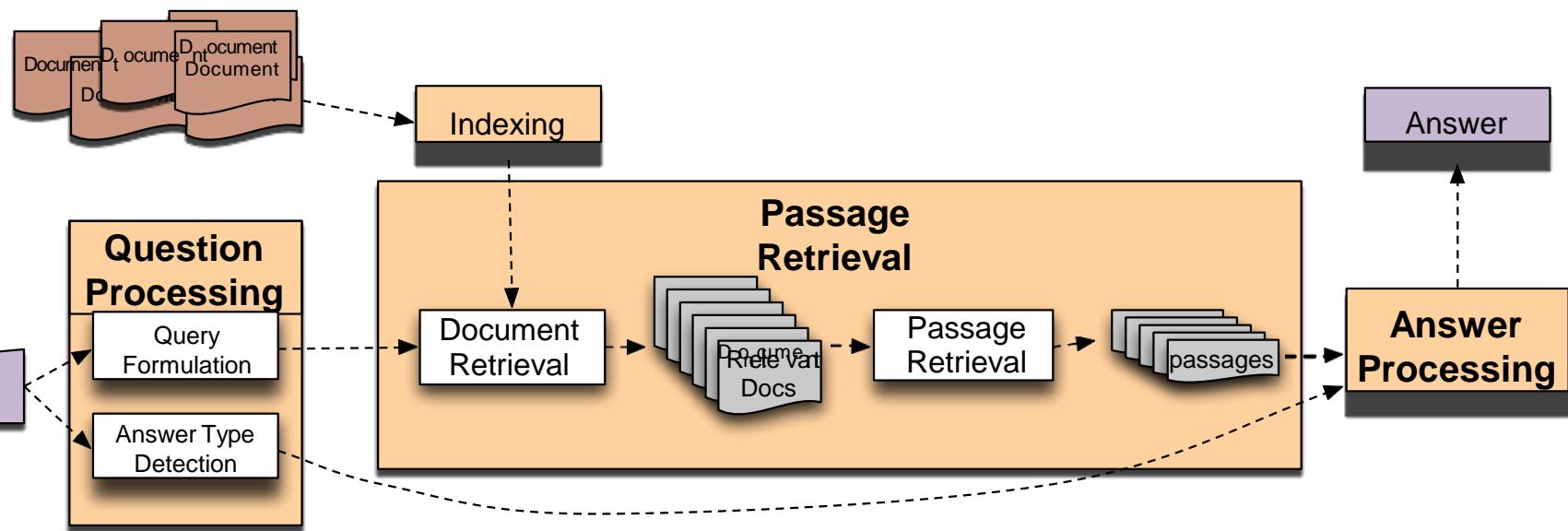
- Build a semantic representation of the query
 - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
 - Geospatial databases
 - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
 - Restaurant review sources and reservation services
 - Scientific databases

Hybrid approaches (IBM Watson)



- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
 - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
 - Geospatial databases
 - Temporal reasoning
 - Taxonomical classification

Factoid Q/A



Question Processing

Things to extract from the question

- Answer Type Detection
 - Decide the **named entity type** (person, place) of the answer
- Query Formulation
 - Choose **query keywords** for the IR system
- Question Type classification
 - Is this a definition question, a math question, a list question?
- Focus Detection
 - Find the question words that are replaced by the answer
- Relation Extraction
 - Find relations between entities in the question

Question Processing

They're the two states you could be reentering if you're crossing Florida's northern border

- Answer Type: US state
- Query: **two states, border, Florida, north**
- Focus: **the two states**
- Relations: **borders(Florida, ?x, north)**

Answer Type Detection: Named Entities



- *Who founded Virgin Airlines?*
 - PERSON
- *What Canadian city has the largest population?*
 - CITY.

Answer Type Taxonomy

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

- 6 coarse classes
 - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
 - LOCATION: city, country, mountain...
 - HUMAN: group, individual, title, description
 - ENTITY: animal, body, color, currency...

Part of Li & Roth's Answer Type Taxonomy

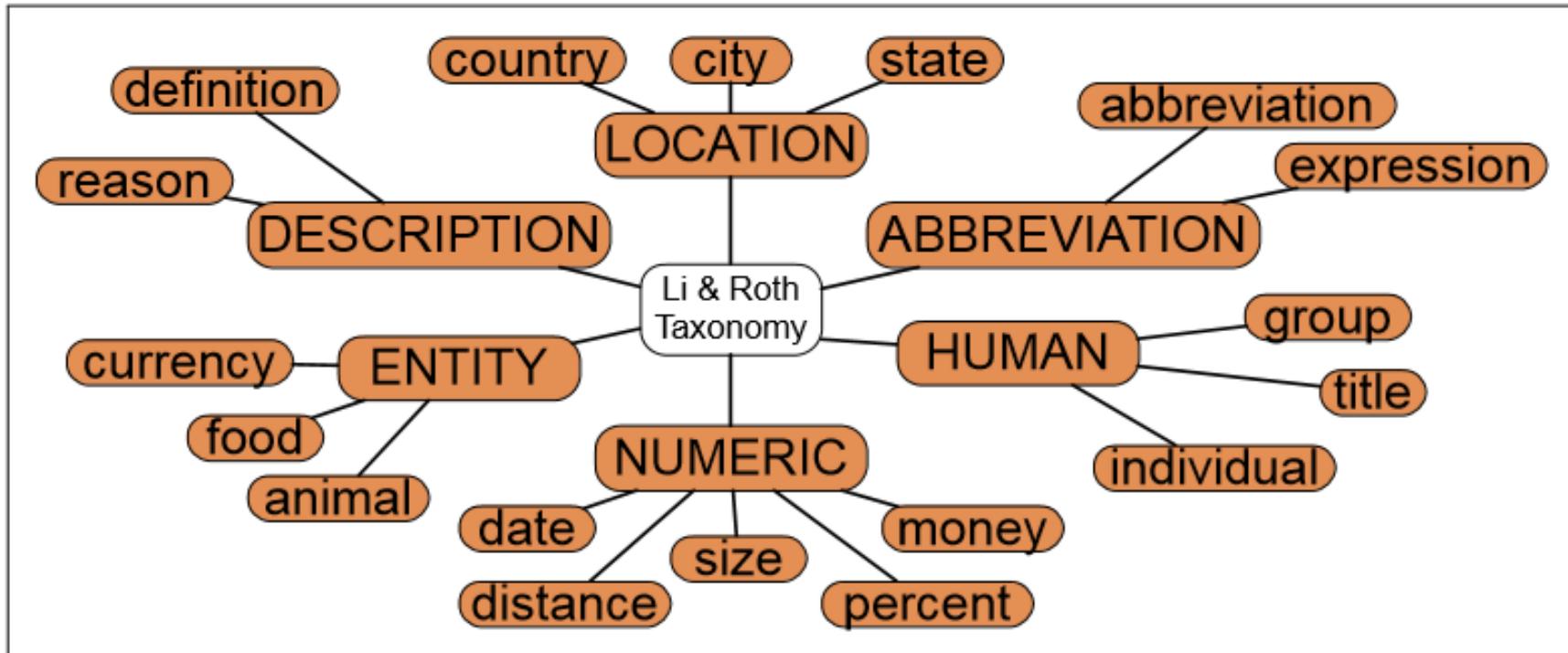


Figure 25.3 A subset of the Li and Roth (2005) answer types.

Answer Types



ENTITY

animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

More Answer Types

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

Answer types in Jeopardy

Ferrucci et al. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. Fall 2010. 59–79.

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types

he, country, city, man, film, state, she, author, group, here, company, president, capital, star, novel, character, woman, river, island, king, song, part, series, sport, singer, actor, play, team, show, actress, animal, presidential, composer, musical, nation, book, title, leader, game

Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

Answer Type Detection

- Regular expression--based rules can get some cases:
 - Who {is|was|are|were} PERSON
 - PERSON (YEAR – YEAR)
- Other rules use the **question headword**:
(the headword of the first noun phrase after the wh-word)
 - Which **city** in China has the largest number of foreign financial companies?
 - What is the state **flower** of California?

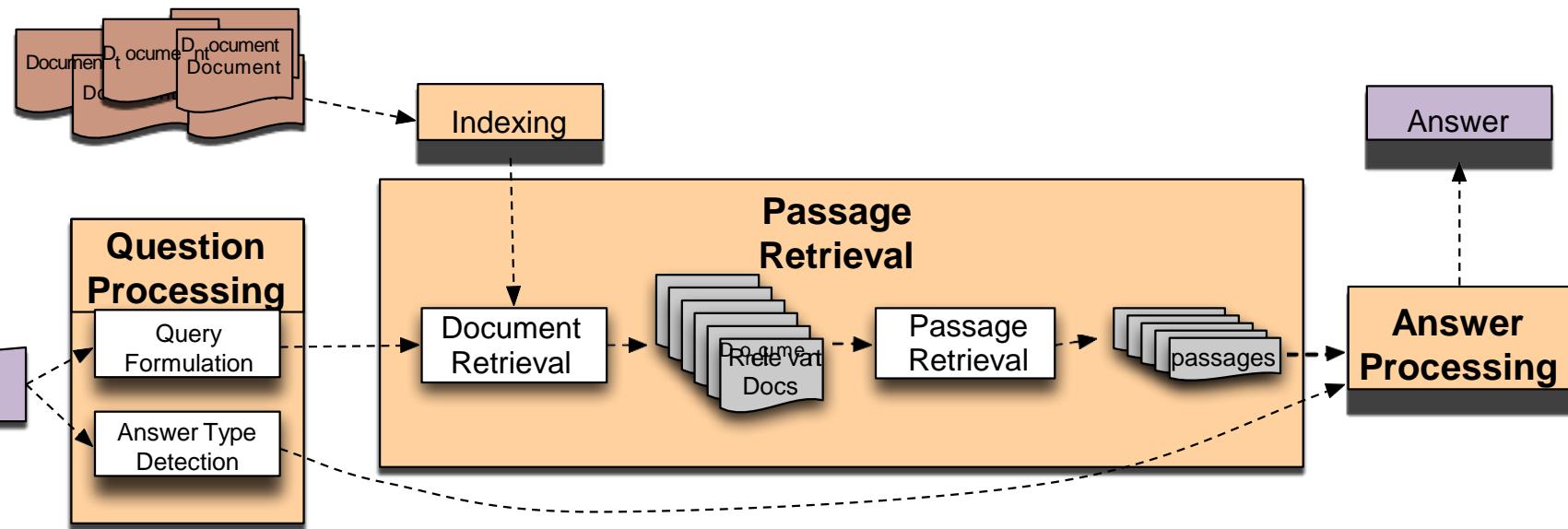
Answer Type Detection

- Most often, we treat the problem as machine learning classification
 - **Define** a taxonomy of question types
 - **Annotate** training data for each question type
 - **Train** classifiers for each question class using a rich set of features.
 - features include those hand-written rules!

Features for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

Factoid Q/A



Keyword Selection Algorithm



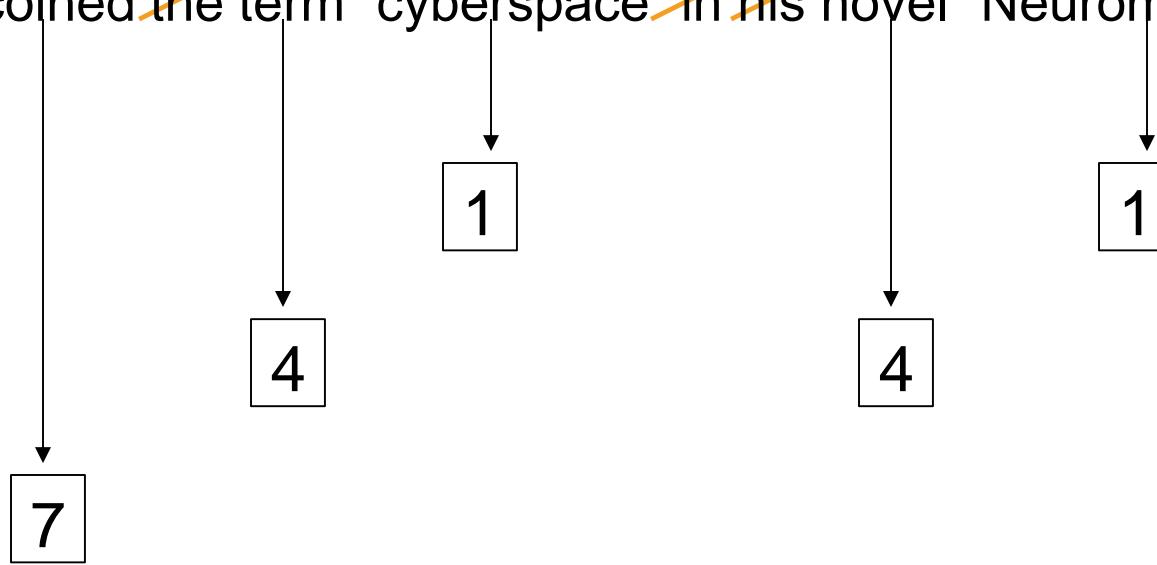
Dan Moldovan, Sanda Harabagiu, Marius Paca, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. 1999. Proceedings of TREC-8.

1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words

Choosing keywords from the query

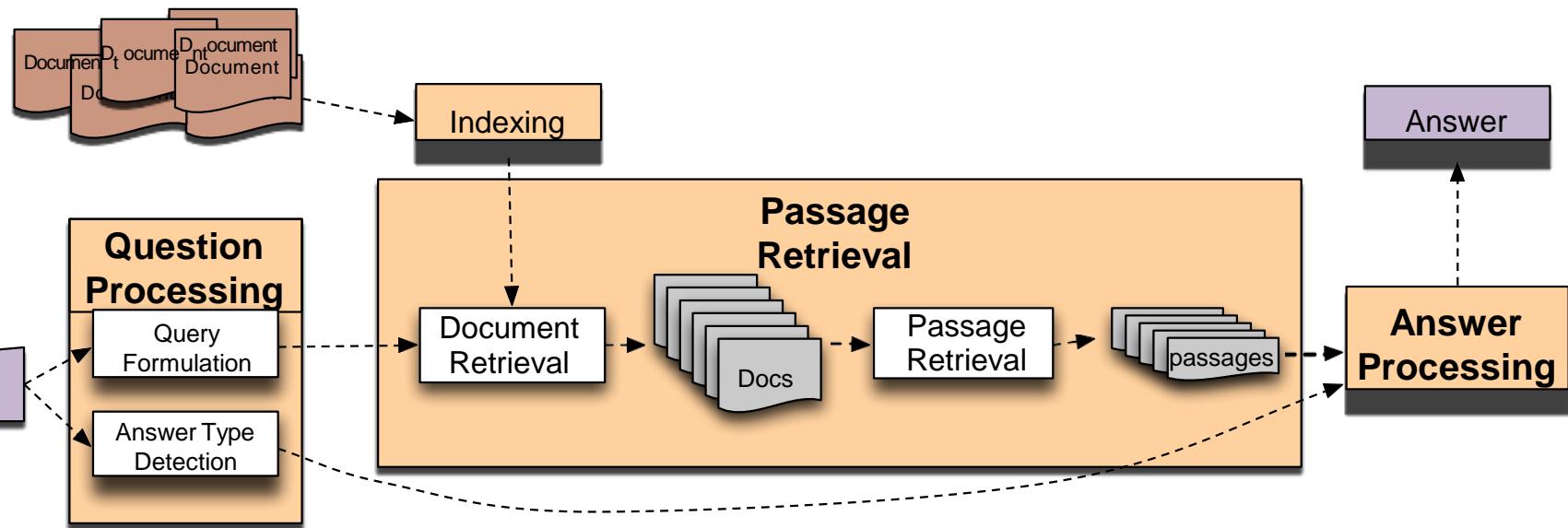
Slide from Mihai Surdeanu

~~Who coined the term “cyberspace” in his novel “Neuromancer”?~~



cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

Factoid Q/A



Passage Retrieval

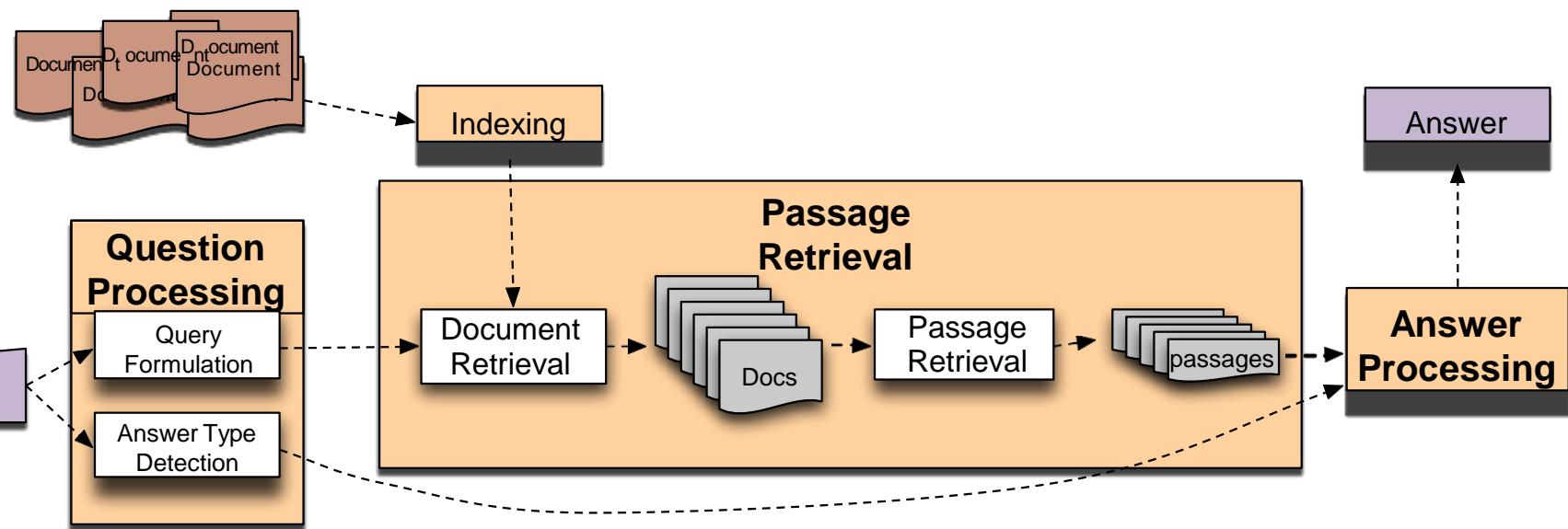
- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
 - something like paragraphs
- Step 3: Passage ranking
 - Use answer type to help rerank passages

Passage Ranking

Either in rule-based classifiers or with supervised machine learning

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

Factoid Q/A



Answer Extraction

- Run an answer-type named-entity tagger on the passages
 - Each answer type requires a named-entity tagger that detects it
 - If answer type is CITY, tagger has to tag CITY
 - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
 - Who was the first prime minister of India (**PERSON**)
The **first** was Jawaharlal Nehru of the **Indian** National Congress party, who was sworn in on 15 August 1947, when **India** gained independence from the British Raj
 - How tall is Mt. Everest? (**LENGTH**)
The official height of Mount Everest is 29035 feet

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**
- Passage:

The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**
- Passage:

The Marie biscuit is named after **Marie Alexandrovna**,
the daughter of **Czar Alexander II of Russia** and wife of
Alfred, the second son of **Queen Victoria** and **Prince
Albert**

Use machine learning: Features for ranking candidate answers

Answer type match: Candidate contains a phrase with the correct answer type.

Pattern match: Regular expression pattern matches the candidate.

Question keywords: # of question keywords in the candidate.

Keyword distance: Distance in words between the candidate and query keywords

Novelty factor: A word in the candidate is not in the query.

Apposition features: The candidate is an appositive to question terms

Punctuation location: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

Sequences of question terms: The length of the longest sequence of question terms that occurs in the candidate answer.

Candidate Answer scoring in IBM Watson



- Each candidate answer gets scores from >50 components
 - (from unstructured text, semi--structured text, triple stores)
 - logical form (parse) match between question and candidate
 - passage source reliability
 - geospatial location
 - California is "southwest of Montana"
 - temporal relationships
 - taxonomic classification

Relation Extraction

- Answers: Databases of Relations
 - born--in("Emma Goldman", "June 27 1869")
 - author--of("Cao Xue Qin", "Dream of the Red Chamber")
 - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions

Whose granddaughter starred in E.T.?

(acted-in ?x "E.T.")

(granddaughter-of ?x ?y)

Temporal Reasoning

- Relation databases
 - (and obituaries, biographical dictionaries, etc.)
- IBM Watson

“In 1594 he took a job as a tax collector in Andalusia”

Candidates:

 - Thoreau is a bad answer (born in 1817)
 - Cervantes is possible (was alive in 1594)

Geospatial knowledge (containment, directionality, borders)

- Beijing is a good answer for "Asian city"
- California is "southwest of Montana"
- geonames.org:

www.geonames.org/search.html?q=palo+alto&country=

[GeoNames Home](#) | [Postal Codes](#) | [Download](#) / [Webservice](#) | [About](#) [login](#)

	palo alto	all countries		
	<input type="text" value="palo alto"/>	<input type="button" value="search"/>	<input type="button" value="show on map"/>	
			[advanced search]	
459 records found for "palo alto"				
Name	Country	Feature class	Latitude	Longitude
1  Palo Alto 	United States , California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30"	W 122° 8' 34"
2  Palo Alto Township	United States , Iowa Jasper County	administrative division elevation 256m	N 41° 38' 15"	W 93° 2' 57"
3  Borough of Palo Alto	United States , Pennsylvania Schuylkill County	administrative division population 1,032, elevation 210m	N 40° 41' 21"	W 76° 10' 2"

Context and Conversation in Virtual Assistants like Siri

- Coreference helps resolve ambiguities
 - U: “Book a table at Il Fornaio at 7:00 with **my mom**”
 - U: “Also send **her** an email reminder”
- Clarification questions:
 - U: “Chicago pizza”
 - S: “Did you mean pizza restaurants in Chicago or Chicago--style pizza?”

Question Answering: IBM's Watson



- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
“AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA”
INSPIRED THIS AUTHOR’S
MOST FAMOUS NOVEL



Bram Stoker

IBM Watson QA

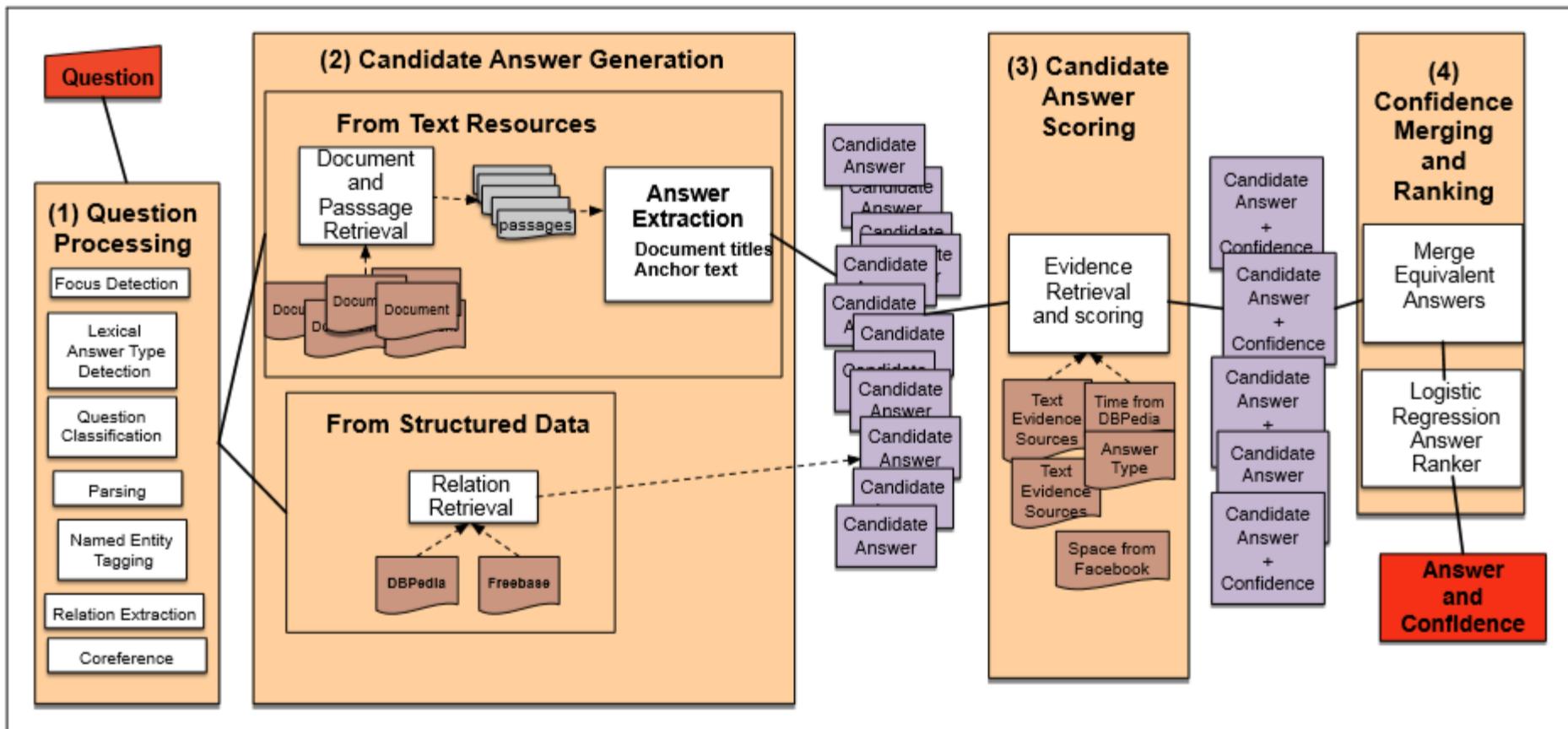


Figure 25.11 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

Stage 1: Question Processing

- Parsing
- Named Entity Tagging
- Relation Extraction
- Focus
- Answer Type
- Question Classification

Named Entity and Parse



Poets and Poetry: He was a bank clerk in the Yukon before
He published “Songs of a Sourdough” in 1907. (he is
linked with clerk) COMPOSITION YEAR

Relations like following are extracted:

authorof(focus,“Songs of a sourdough”) publish
(e1, he, “Songs of a sourdough”) in (e2, e1, 1907)
temporallink(publish(...), 1907)

Named Entity and Parse

THEATRE: A new play based on **this**

Sir Arthur Conan Doyle canine classic

PERSON

opened on the **London** stage in **2007.**

GEO

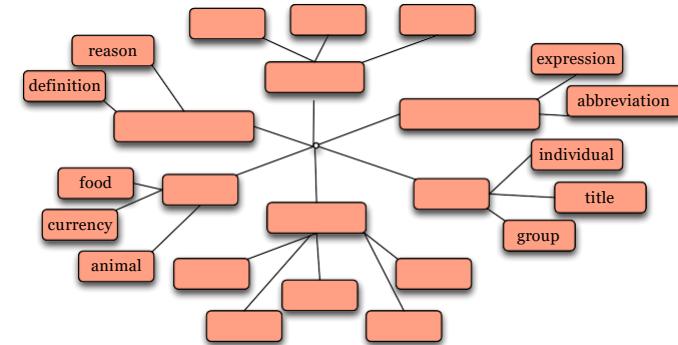
YEAR

Focus extraction

- **Focus:** the part of the question that co-refers with the answer
- Replace it with answer to find a supporting passage.
- Extracted by hand-written rules
 - "Extract any noun phrase with determiner this"
 - “Extracting pronouns *she, he, hers, him,* ”

Lexical Answer Type

- The semantic class of the answer
- But for Jeopardy the TREC answer type taxonomy is insufficient
- DeepQA team investigated 20,000 questions
- 100 named entities only covered <50% of the questions!
- Instead: Extract lots of words: 5,000 for those 20,000 questions



Lexical Answer Type

- Answer types extracted by hand-written rules
 - Syntactic headword of the focus.
 - Words that are coreferent with the focus
 - Jeopardy! category, if refers to compatible entity.

Poets and Poetry: **He** was a bank **clerk** in the Yukon before he published “Songs of a Sourdough” in 1907.

Relation Extraction in DeepQA

- For the most frequent 30 relations:
 - Hand-written regular expressions
 - AuthorOf:
 - Many patterns such as one to deal with:
 - *a Mary Shelley tale, the Saroyan novel, Twain's travel books, a 1984 Tom Clancy thriller*
 - [Author] [Prose]
- For the rest: distant supervision

Stage 2: Candidate Answer Generation



Extracting candidate answers from triple stores

- If we extracted a relation from the question
... he published “Songs of a sourdough”
(author-of ?x “Songs of a sourdough”)
- We just query a triple store
 - Wikipedia infoboxes, DBpedia, FreeBase, etc.
 - born--in(“Emma Goldman”, “June 27 1869”)
 - author--of(“Cao Xue Qin”, “Dream of the Red Chamber”)
 - author--of(“Songs of a sourdough”, “Robert Service”)

Extracting candidate answers from text: get documents/passages



Do standard IR-based QA to get documents

Robert Redford and Paul Newman starred in this depression--era griter flick.

(2.0 Robert Redford) (2.0 Paul Newman) star depression era griter (1.5 flick)

Extracting answers from documents/ passages



- Useful fact: Jeopardy! answers are mostly the title of a Wikipedia document
 - If the document is a Wikipedia article, just take the title
 - If not, extract all noun phrases in the passage that are Wikipedia document titles
 - Or extract all anchor text <a>The Sting

Stage 3: Candidate Answer Scoring



- **Use lots of sources of evidence to score an answer**
 - more than 50 scorers
- **Lexical answer type is a big one**
 - Different in DeepQA than in pure IR factoid QA
 - In pure IR factoid QA, answer type is used to strictly filter answers
 - In DeepQA, answer type is just one of many pieces of evidence

Lexical Answer Type (LAT) for Scoring Candidates



- Given:
 - candidate answer & lexical answer type
- Return a score: can answer can be a subclass of this answer type?
- Candidate: “*difficulty swallowing*” & LAT “*condition*”
 1. Check DBpedia, WordNet, etc
 - *difficulty swallowing* → Dbpedia *Dysphagia* → WordNet *Dysphagia*
 - *condition* → WordNet *Condition*
 2. Check if “Dysphagia” IS-A “Condition” in WordNet
 - [Wordnet for dysphagia](#)

Relations for scoring

- Q: This hockey defenseman ended his career on June 5, 2008
- **Passage:** On June 5, 2008, Wesley announced his retirement after his 20th NHL season

- Question and passage have very few keywords in common
- But both have the Dbpedia relation ActiveYearsEndDate()

Temporal Reasoning for Scoring Candidates



- Relation databases
 - (and obituaries, biographical dictionaries, etc.)
- IBM Watson

"In 1594 he took a job as a tax collector in Andalusia"

Candidates:

 - Thoreau is a bad answer (born in 1817)
 - Cervantes is possible (was alive in 1594)

Geospatial knowledge (containment, directionality, borders)

- Beijing is a good answer for "Asian city"
- California is "southwest of Montana"
- geonames.org:



The screenshot shows a web browser displaying the GeoNames search results for "palo alto". The search bar contains "palo alto" and the dropdown menu shows "all countries". Below the search bar are buttons for "search", "show on map", and "[advanced search]". The page title is "GeoNames Home | Postal Codes | Download / Webservice | About" and there is a "login" link. The main content area displays a table with 459 records found for "palo alto". The table has columns: Name, Country, Feature class, Latitude, and Longitude. The first result is "Palo Alto" in the United States, California, Santa Clara County, a populated place with a population of 64,403, located at N 37° 26' 30" and W 122° 8' 34". The second result is "Palo Alto Township" in the United States, Iowa, Jasper County, an administrative division with an elevation of 256m, located at N 41° 38' 15" and W 93° 2' 57". The third result is "Borough of Palo Alto" in the United States, Pennsylvania, Schuylkill County, an administrative division with a population of 1,032, located at N 40° 41' 21" and W 76° 10' 2".

Name	Country	Feature class	Latitude	Longitude
1  Palo Alto  Palo Al'to, Palo Alto, pa luo ao duo, paroaruto, Пало Алто, Пало Альто, פאלו אלטו, パロアルト, 帕羅奧多	United States , California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30"	W 122° 8' 34"
2  Palo Alto Township Palo Alto Township	United States , Iowa Jasper County	administrative division elevation 256m	N 41° 38' 15"	W 93° 2' 57"
3  Borough of Palo Alto	United States , Pennsylvania Schuylkill County	administrative division population 1,032, elevation 210m	N 40° 41' 21"	W 76° 10' 2"

Stage 4: Answer Merging and Scoring



- Now we have a list candidate answers each with a score vector
 - J.F.K [.5 .4 1.2 33 .35 ...]
 - John F. Kennedy [.2 .56 5.3 2 ...]
- Merge equivalent answers: *J.F.K.* and *John F. Kennedy*
 - Use Wikipedia dictionaries that list synonyms:
 - *JFK, John F. Kennedy, John Fitzgerald Kennedy, Senator John F. Kennedy, President Kennedy, Jack Kennedy*
 - Use stemming and other morphology

Stage 4: Answer Scoring

- Build a classifier to take answers and a score vector and assign a probability
- Train on datasets of hand-labeled correct and incorrect answers.

QA in Deep Learning

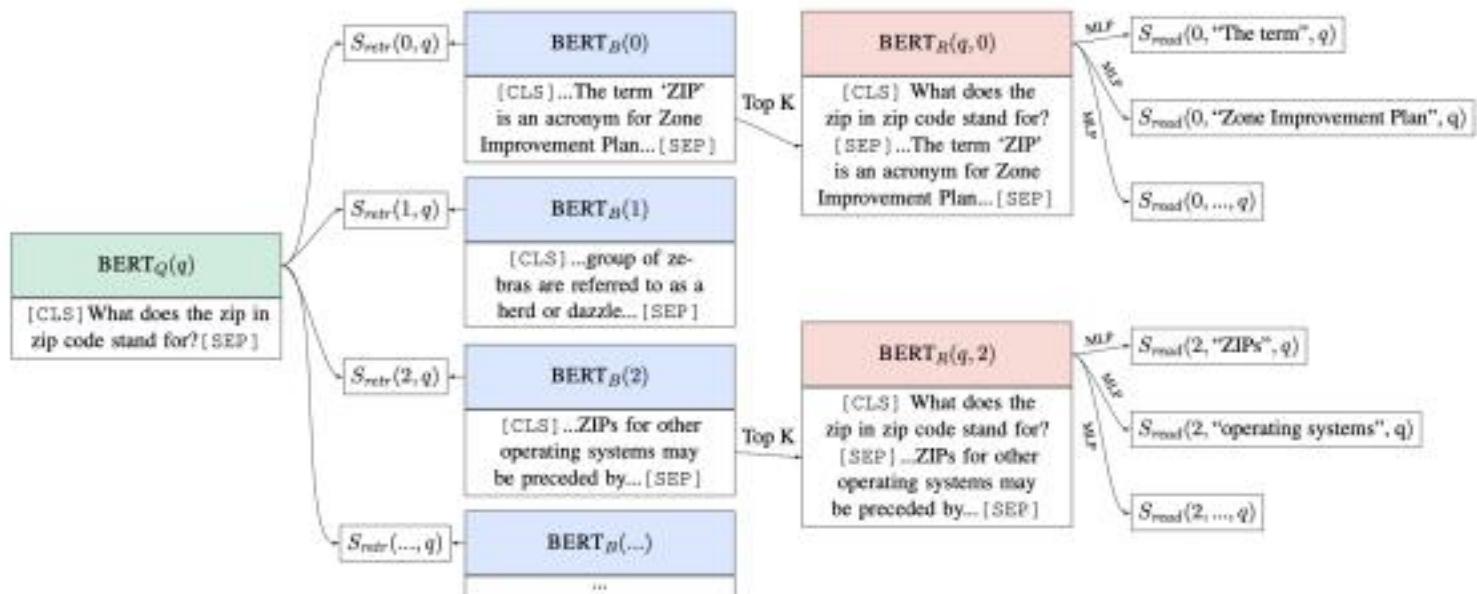


Image credit: (Lee et al., 2019)

Almost all the state-of-the-art question answering systems are built on top of end-to-end training and pre-trained language models (e.g., BERT)!

QA training datasets

Stanford question answering dataset (SQuAD)

- 100k annotated (passage, question, answer) triples

Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!
- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.

This is a limitation— not all the questions can be answered in this way!
- SQuAD was for years the most popular reading comprehension dataset; it is “almost solved” today (though the underlying task is not,) and the state-of-the-art exceeds the estimated human performance.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

QA training datasets

Other question answering datasets

- TriviaQA: Questions and answers by trivia enthusiasts. Independently collected web paragraphs that contain the answer and seem to discuss question, but no human verification that paragraph supports answer to question
- Natural Questions: Question drawn from frequently asked Google search questions. Answers from Wikipedia paragraphs. Answer can be substring, yes, no, or NOT_PRESENT. Verified by human annotation.
- HotpotQA. Constructed questions to be answered from the whole of Wikipedia which involve getting information from two pages to answer a multistep query:
Q: Which novel by the author of “Armada” will be adapted as a feature film by Steven Spielberg? A: *Ready Player One*

QA in Deep Learning

How can we build a model to solve SQuAD?

(We are going to use **passage**, **paragraph** and **context**, as well as **question** and **query** interchangeably)

- Problem formulation
 - Input: $C = (c_1, c_2, \dots, c_N), Q = (q_1, q_2, \dots, q_M), c_i, q_i \in V$ $N \sim 100, M \sim 15$
 - Output: $1 \leq \text{start} \leq \text{end} \leq N$ answer is a span in the passage
- A family of LSTM-based models with attention (2016–2018)

Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), BiDAF (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..
- Fine-tuning BERT-like models for reading comprehension (2019+)

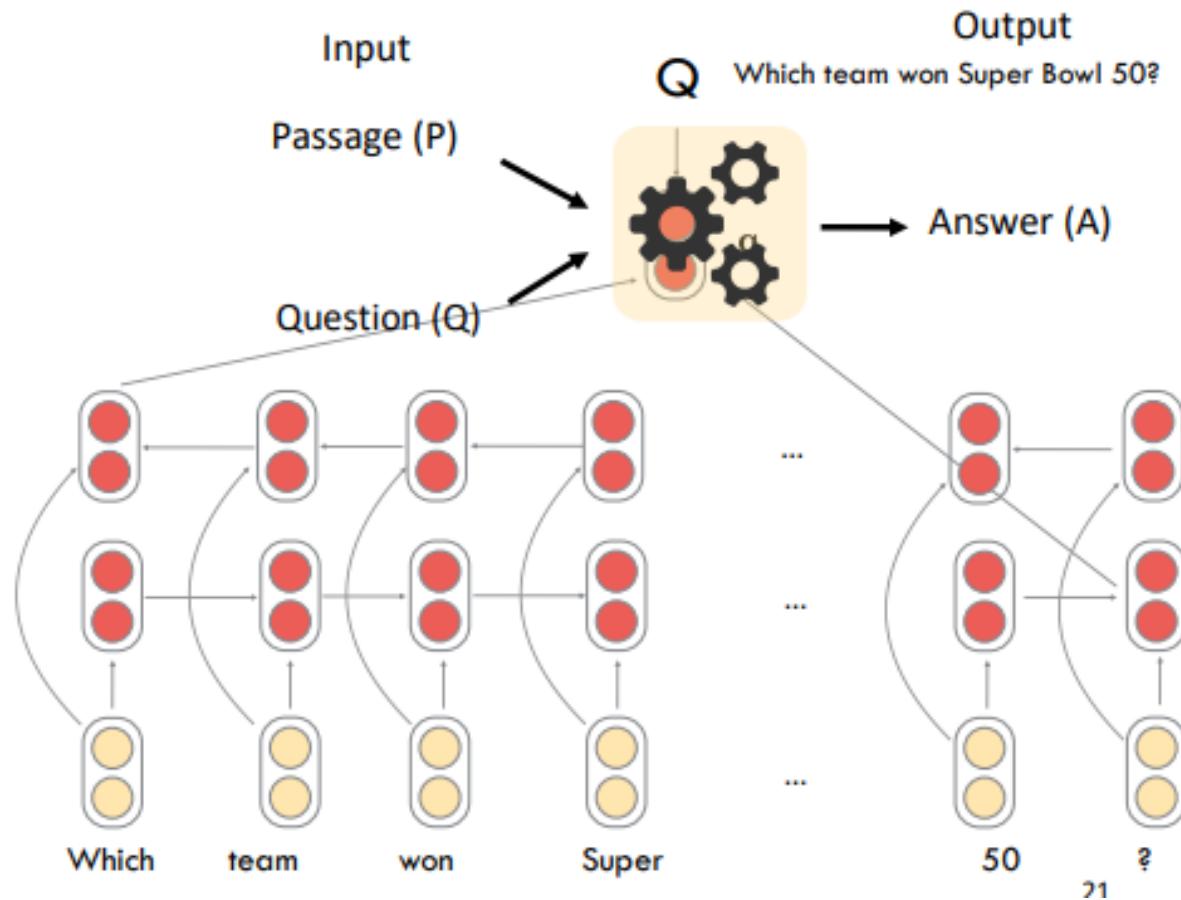
QA in Deep Learning

- Transfer learning is the application of knowledge learned while solve one problem on other similar problems.
- Latest deep learning based word embedding's such as Bidirectional Encoder Representation from Transformers (BERT) *enable pre-trained Question Answer models trained on corpus from one domain to easily answer questions from another domain.*
- This makes is easy to introduce support for Question Answering in newer applications using pre-trained models.

Word embedding of Question

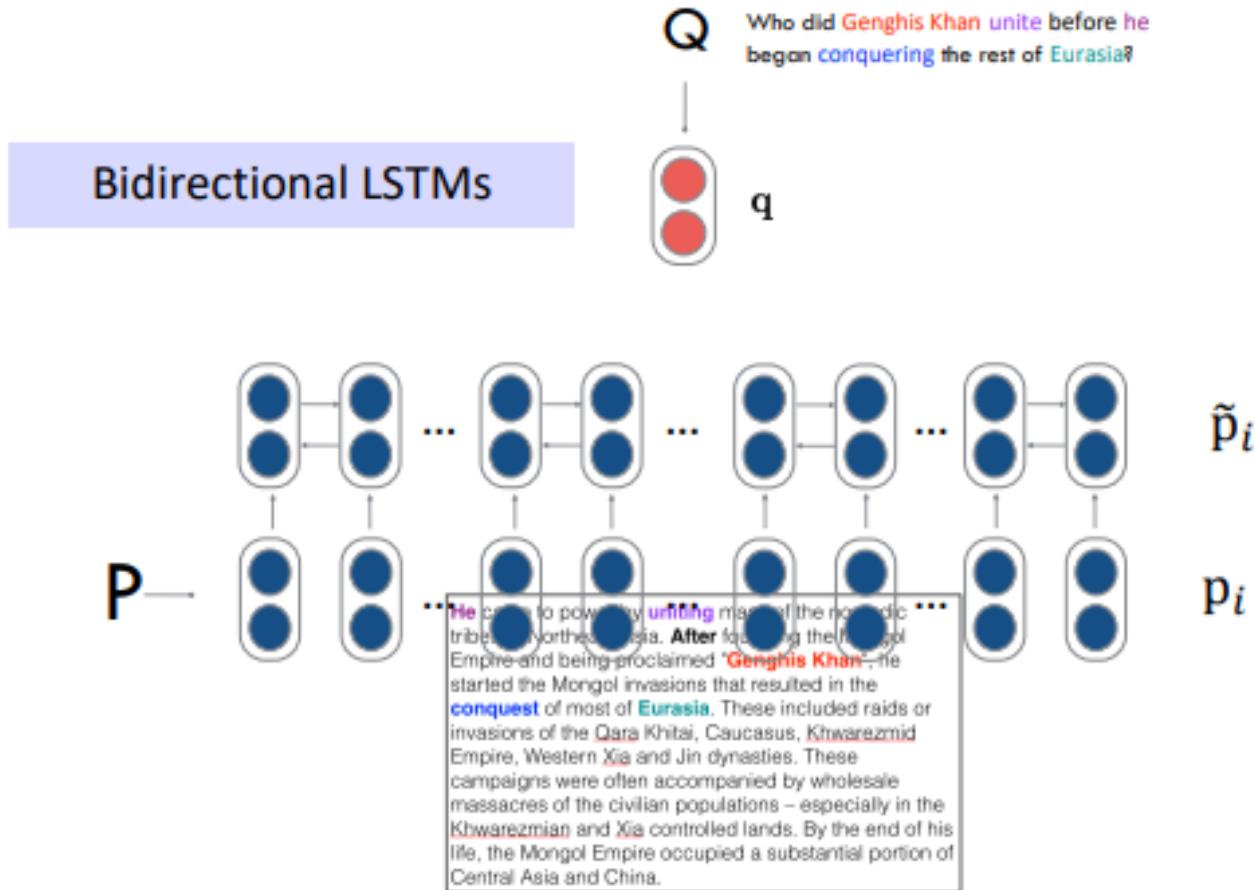
The Stanford Attentive Reader

(Use RNN in both directions - BiLSTM)



Word embedding of passage

Stanford Attentive Reader



Word embedding of passage

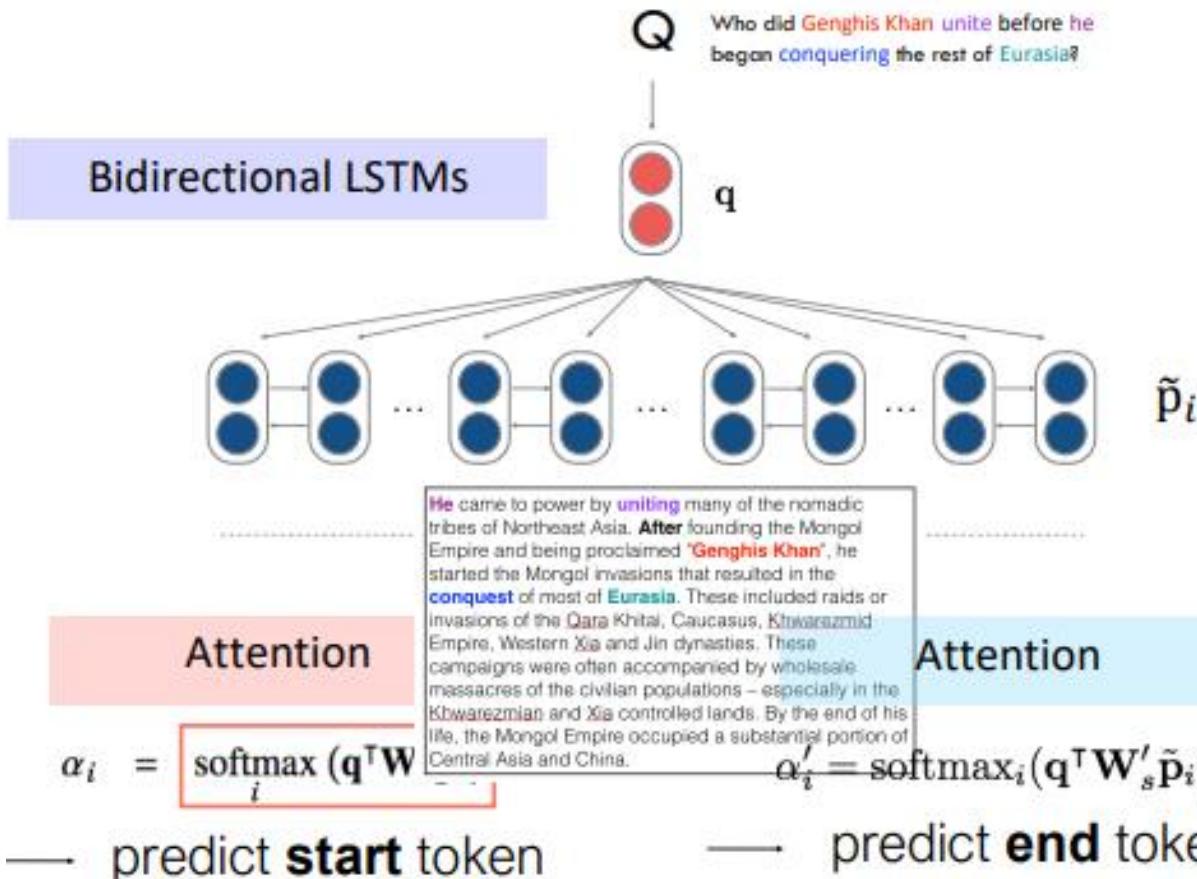
p_i : Vector representation of each token in passage

Made from concatenation of

- Word embedding (GloVe 300d)
- Linguistic features: POS & NER tags, one-hot encoded
- Term frequency (unigram probability)
- Exact match: whether the word appears in the question
- 3 binary features: exact, uncased, lemma
- Aligned question embedding (“car” vs “vehicle”)

Attention Mechanism to predict start and end token of answer

Stanford Attentive Reader



Deep Learning architecture for QA

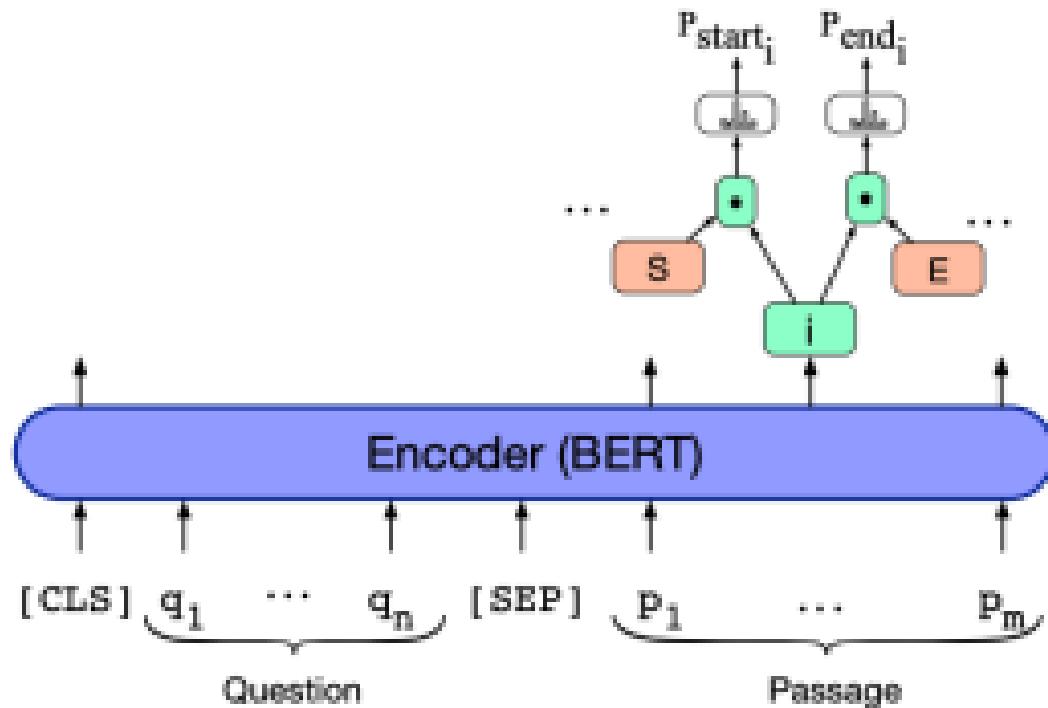
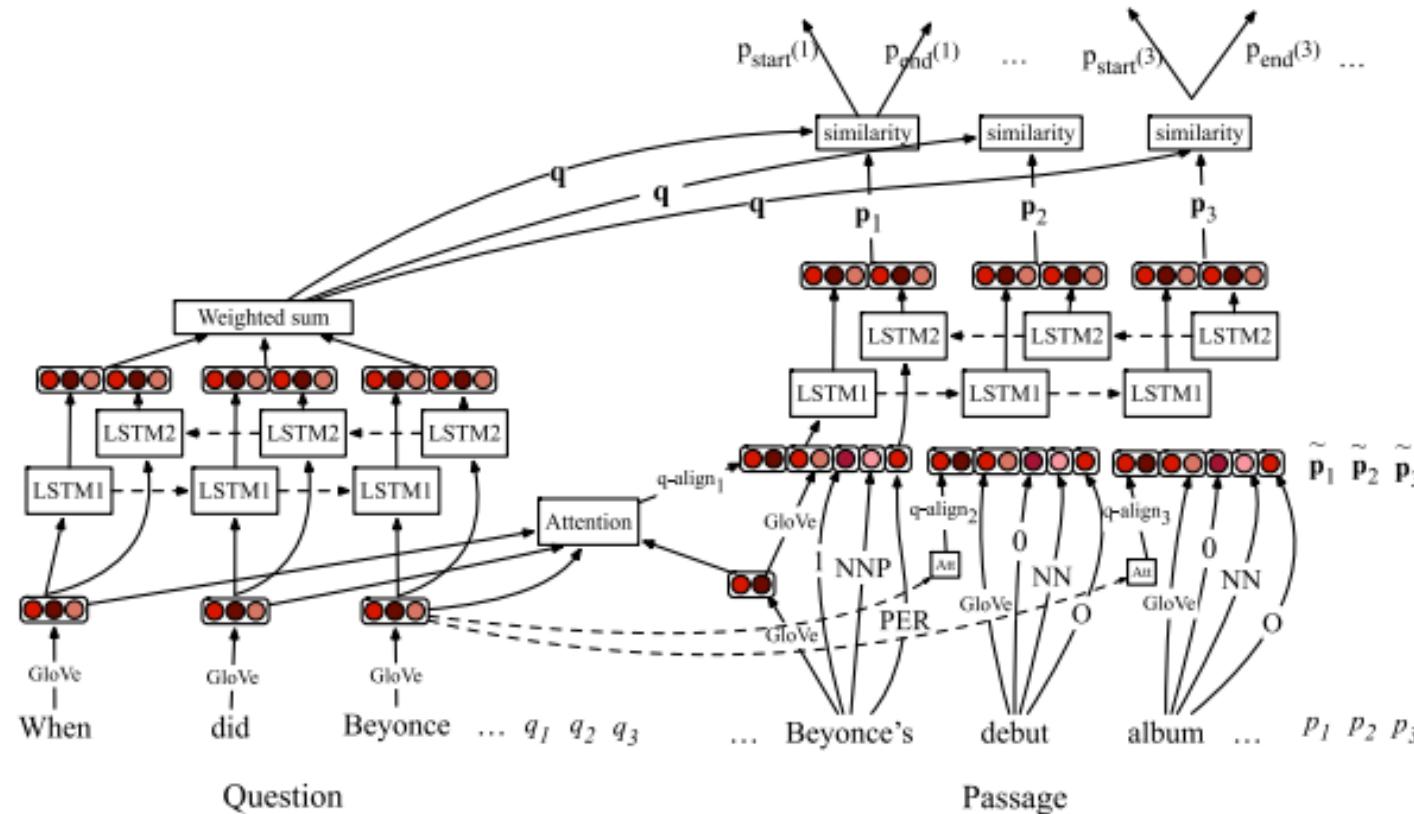


Image credit: J & M, edition 3

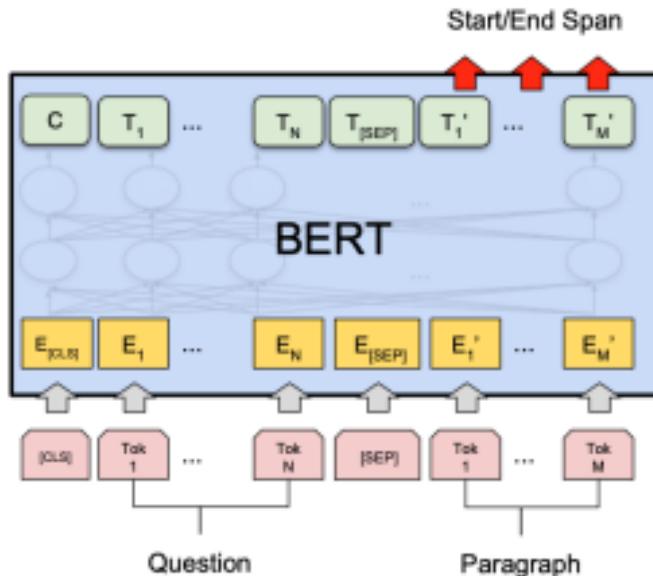
Deep Learning architecture for QA



Training objective: $\mathcal{L} = -\sum \log P^{(\text{start})}(a_{\text{start}}) - \sum \log P^{(\text{end})}(a_{\text{end}})$

Deep Learning architecture for QA

BERT for reading comprehension



Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B



Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

Image credit: <https://mccormickml.com/>

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

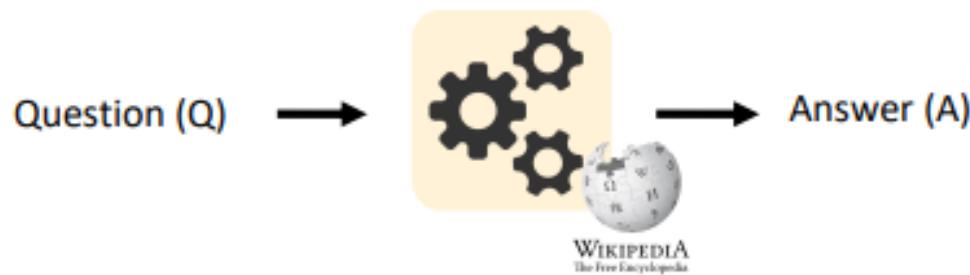
$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^\top \mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^\top \mathbf{h}_i)$$

where \mathbf{h}_i is the hidden vector of C_i , returned by BERT

Open Domain QA

Open-domain question answering

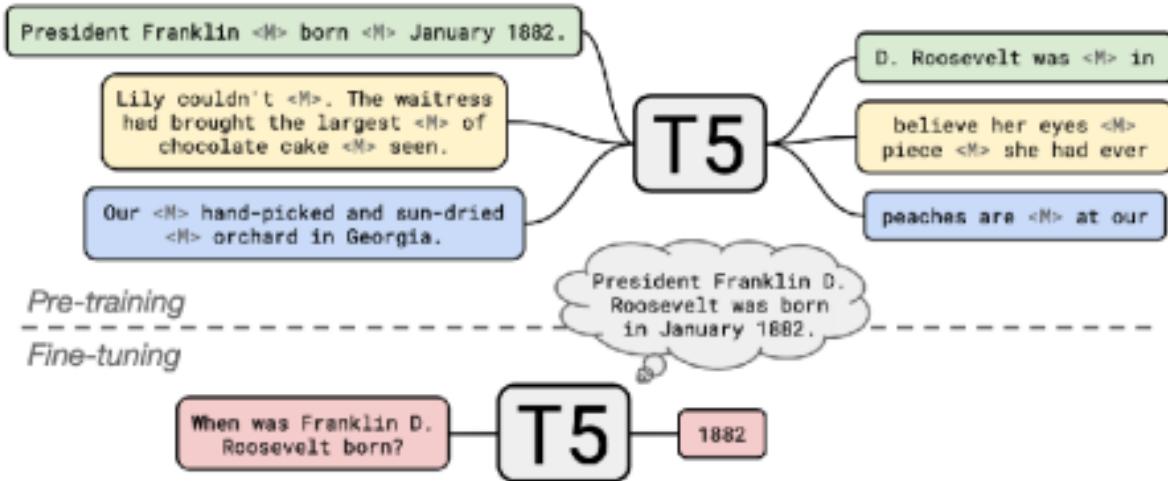


- Different from reading comprehension, we don't assume a given passage.
- Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don't know where the answer is located, and the goal is to return the answer for any open-domain questions.
- Much more challenging and a more practical problem!

T5 LLM for QA

Large language models can do open-domain QA well

- ... without an explicit retriever stage



Text-to-Text Transfer Transformer (T5)

- The architecture of T5 model is almost the same as the BERT.
- Both the encoder and decoder consist of 12 blocks.
- This model has 220 million parameters.
- Only a few changes have been made to the architecture like they have removed the Layer norm bias and placed the layer normalization outside the residual path.
- There is a different position embedding scheme used in T5.

Challenges of Deep learning approaches



Every manager is an employee.
Rose is a manager.
Rose is an employee?

Reasoning

Google was founded by Larry Page. Sergey Brin was a co-founder.
Who founded Google?

Answer spanning multiple parts of the document

Who are the founders of Google and Facebook?

Multiple Questions

Challenges of Information Retrieval Based Systems Developed Using Deep Learning Based Question Answer models

Challenges of Deep learning approaches



- Incapable of answering questions that require reasoning.
- Deep learning based models for Question Answering take the input passage and questions and output the start and end position in the passage that contains the answer.
- Consequently they can not answer questions whose answer is spread across the document.
- Can not answer questions that have multiple sub questions whose answers are spread across the document.

Common Evaluation Metrics

1. Accuracy (does answer match gold-labeled answer?)

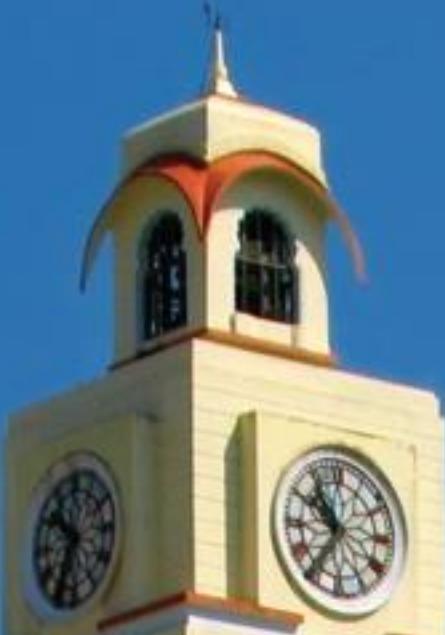
2. Mean Reciprocal Rank

- For each query return a ranked list of M candidate answers.
- Query score is $1/\text{Rank}$ of the first correct answer
 - *If first answer is correct: 1*
 - *else if second answer is correct: $\frac{1}{2}$*
 - *else if third answer is correct: $\frac{1}{3}$, etc.*
 - *Score is 0 if none of the M answers are correct*
- Take the mean over all N queries

$$MRR = \frac{\sum_{i=1}^N \frac{1}{rank_i}}{N}$$

References

- Speech and Language processing: An introduction to Natural Language Processing, Computational Linguistics and speech Recognition by Daniel Jurafsky and James H. Martin[3rd edition] - Chapter 25
- <https://medium.com/@akashgoel12/question-answering-system-on-corona-approach-01-6ef9799695cb>
- <https://www.machinelearningplus.com/nlp/chatbot-with-rasa-and-spacy/>
- <https://analyticsindiamag.com/10-question-answering-datasets-to-build-robust-chatbot-systems/>
- <https://github.com/ElizaLo/Question-Answering-based-on-SQuAD>
- <https://intersog.com/blog/the-basics-of-qa-systems-from-a-single-function-to-a-pre-trained-nlp-model-using-python/>
- <https://paperswithcode.com/task/question-answering/latest>
- NN approach
- https://www.youtube.com/watch?v=NcqfHa0_YmU
- https://www.youtube.com/watch?v=3XiJrn_8F9Q
- <https://towardsdatascience.com/implementing-question-answering-networks-with-cnns-5ae5f08e312b>



Natural Language Processing Applications



BITS Pilani
Pilani Campus

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 5: Dialogue Systems and Chatbots

Date – 24th Dec 2023

Time – 1.40 pm to 3.40 pm

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Dan Jurafsky and many others who made their course materials freely available online.

Session Content

- Properties of Human Conversation
 - Chabot's
 - GUS: Simple Frame-based Dialogue Systems
 - The Dialogue-State Architecture
 - Evaluating Dialogue Systems
 - Dialogue System Design
-

AI Assistants

Notification Assistant



Hi there - just a friendly reminder that your insurance policy expires in a month. Make sure to renew it in our member portal.

FAQ Assistant



I need to renew my renters insurance. How much will it be?



You can calculate your renewal price on our website here:
xyz.com/renew

Contextual Assistant



I need to renew my renters insurance. How much will it be?



I'd be happy to check for you. Firstly, are you still living in the same apartment?

Yes



Great - so just confirming it's 980 sq ft?

Yes



Thanks! Your new rate from September 1st onwards would be \$10 / month.



Would you like me to renew your policy for you right now?

Sure



Great. I've sent you a confirmation to your email.

AI Assistants

Personalized Assistant

- Assistant knows you much more in detail
- Quickly checks a few final things before giving you a quote tailored to your actual situation.



I can see your details are almost the same, except now you might want coverage for your new laptop. Additional coverage is only \$4 a month more for full coverage.
Sound ok?

Sounds good!



Autonomous Organization of Assistants

- Group of AI assistants that know every customer personally
- Eventually run large parts of company operations—from lead generation over marketing, sales, HR, or finance



Conversational Agents

AKA Dialog Agents

- Phone-based Personal Assistants
 - SIRI (Apple), Alexa (Amazon),
 - Cortana (Microsoft), Google Assistant
 - Talking to your car, Pay bills...
 - Sales, Marketing, Insurance....
 - Clinical uses for mental health, Nurses, Dr. Bot
 - Lawyer bots
 - Chatting for fun, Mr. FriendBot
-

Context and Conversation in Virtual Assistants like Siri

- Coreference helps resolve ambiguities
 - U: “Book a table at Il Fornaio at 7:00 with **my mom**”
 - U: “Also send **her** an email reminder”
- Clarification questions:
 - U: “Chicago pizza”
 - S: “Did you mean pizza restaurants in Chicago or Chicago--style pizza?”

Conversation Characteristics

Turns

- dialogue is a sequence of turns
- C1: I need to travel in May.
A1: And, what day in May did you want to travel?
C2: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A2: And you're flying into what city? C3: Seattle.

Speech Acts of Dialog Acts:

- Constitutives: answering, claiming, confirming, denying, disagreeing, stating)
- Directives: to do something (advising, asking, forbidding, inviting, ordering, requesting)
- Commissives: committing the speaker to some future course of action (promising, planning, vowing, betting, opposing)
- Acknowledgments: apologizing, greeting, thanking, accepting an acknowledgment

Challenges of Dialog Systems

Grounding:

- acknowledging that the hearer has understood the speaker

Dialog Structure:

- QUESTIONS set up an expectation for an ANSWER. PROPOSALS are followed by ACCEPTANCE (or REJECTION). COMPLIMENTS (“Nice jacket!”) often give rise to adjacency pair DOWNPLAYERS (“Oh, this old thing?”). These pairs, called adjacency pairs are composed of a first pair part and a second pair part

Clarification/correction question

- Which can form a sub dialogue between a REQUEST and a RESPONSE
- Ex: User: What do you have going to UNKNOWN WORD on the 5th?
System: Let's see, going where on the 5th?

Inference:

- A1: And, what day in May did you want to travel? C2: OK uh I need to be there for a meeting that's from the 12th to the 15th. Notice that the client does not in fact answer the agent's question. The client merely mentions a meeting at a certain time

Two classes of systems

1. Chatbots/ Chit-Chats
2. (Goal-based) Task oriented
Dialog agents
 - *SIRI, interfaces to cars, robots,*
 - *booking flights or restaurants*

The word "chatbots" is sometimes used in the popular press for both. We'll use it only for #1.

Part I: Chatbots!

- ELIZA (1966)
- PARRY (1968)
 - The first system to pass the Turing test!!!!
- ALICE
- CLEVER
- Microsoft Little Bing

Chatbot Architectures

Rule-based

1. Pattern-action rules (Eliza)
+ a mental model (Parry)

Corpus-based (from large chat corpus)

2. Information Retrieval
3. Neural network encoder-decoder

Chatbots

- Sometimes for fun
- But originally designed as a testbed for clinical therapy
(And still used that way!)

Eliza: Weizenbaum (1966)

Dialog in Green are by Eliza system.

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

...

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?

My father

YOUR FATHER

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

You are not very aggressive but I think you don't want me to notice that.

WHAT MAKES YOU THINK I AM NOT AGGRESSIVE

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU

My father is afraid of everybody.

WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

That trick: be a Rogerian psychologist

- Draw the patient out by reflecting patient's statements back at them
- Rare type of conversation in which one can “assume the pose of knowing almost nothing of the real world”

Rogerian psychologist

Patient: "I went for a long boat ride"

Psychiatrist: "Tell me about boats"

- You don't assume she didn't know what a boat is
- You assume she had some conversational goal
- Chatbots trying to pass the Turing test usually choose some such domain

Eliza pattern/transform rules

(0 YOU 0 ME) *[pattern]*

→

(WHAT MAKES YOU THINK I 3 YOU) *[transform]*

0 means Kleene *

The 3 is the constituent 0 in pattern

You hate(3) me

WHAT MAKES YOU THINK I HATE YOU

Memory

```
(MEMORY MY  
  (0 YOUR 0 = LETS DISCUSS FURTHER WHY YOUR 3)  
  (0 YOUR 0 = EARLIER YOU SAID YOUR 3)
```

Whenever “MY” is highest keyword

- Randomly select a transform on the MEMORY list
- Apply to sentence
- Store on a stack

Later, if no keyword matches a sentence

- Return the top of the MEMORY queue instead

A hierarchical model of discourse

Other Eliza stuff

Rules can refer to classes of words

Family = mother, father, brother, sister

NOUN = ...

Don't reuse transforms in the same conversation

- Whenever we use a transform associated with a pattern
- We increment a counter for that rule
- So the next time we use the next ranked transform

Some basic transforms happen during input processing

- I -> YOU
- YOU -> I

function ELIZA GENERATOR(*user sentence*) **returns** *response*

Find the word *w* in *sentence* that has the highest keyword rank

if *w* exists

 Choose the highest ranked rule *r* for *w* that matches *sentence*

response \leftarrow Apply the transform in *r* to *sentence*

if *w* = 'my'

future \leftarrow Apply a transformation from the 'memory' rule list to *sentence*

 Push *future* onto memory stack

else (no keyword applies)

either

response \leftarrow Apply the transform for the NONE keyword to *sentence*

or

response \leftarrow Pop the top response from the memory stack

return(*response*)

Some implications

- People became deeply emotionally involved with the program
- Weizenbaum tells the story of his secretary who would ask Weizenbaum to leave the room when she talked with ELIZA
- When he suggested that he might want to store all the ELIZA conversations for later analysis, people immediately pointed out the privacy implications
 - Suggesting that they were having quite private conversations with ELIZA

Parry

- Colby 1971 at Stanford
- Same pattern-response structure as Eliza
- But a much richer:
 - control structure
 - language understanding capabilities
 - mental model: Parry has affective variables
 - Anger, Fear, Mistrust
 - “If Anger level is high, respond with hostility”
- The first system to pass the Turing test (in 1971)
 - Psychiatrists couldn't distinguish interviews with PARRY from (text transcripts of) interviews with real paranoids

Flare concepts

- List of concepts related to Mafia
- An ordered graph designed to lead interviewer to topic
horses→ horse racing→gambling→bookies→underworld→Mafia
- The mention of a new flare topic by interviewer causes a rise in Fear
- Flare topics cause Parry to give preset responses to that flare

Each sentence is mapped into a conceptualization

- A predication on a conceptual object
- A predication on a relation between two objects
- A predication on an attribute:

What is your work?

What sort of work do you do?

Where do you work?

→ (your work?)

What do you do for a living?

What is your job?

Do you have a job?

What is your occupation

- Complex Pattern/transform rules

- Different predicates (fear, afraid of)
- Ordering (You are afraid of me = I frighten you)

A Parry conversation

(1) How are you today?

OK.

(2) What is your name?

MY NAME IS FRANK SMITH.

(3) How old are you?

TWENTY-EIGHT.

(4) Are you married?

NO.

(5) What sort of work do you do?

I WORK IN THE POST OFFICE.

(6) Why are you in the hospital?

I SHOULDN'T BE HERE

(7) Who brought you here?

THE POLICE

[FEAR rises slightly because police linked in graph to MAFIA]

(9) What trouble did you have with the police

COPS DON'T DO THEIR JOB. *[police' acts as a flare]*

(10) What do they do instead?

COPS ARREST THE WRONG PEOPLE.

(11) They arrest the right ones sometimes

THE LAW SHOULD GET THOSE ITALIAN CROOKS.

(12) Do you know any Italian crooks?

I TRY TO AVOID THE UNDERWORLD.

(13) Are you afraid of the underworld?

THE MAFIA ARE THE WORST GANGSTERS.

(15) Are you connected with the Mafia?

YOU KNOW, THEY KNOW ME.

(16) How do they know you?

THEY SPY ON ME.

(20) ARE YOU SURE OF ALL THIS ?

YOU DON'T BELIEVE ME ? *[FEAR, ANGER, MISTRUST GO UP]*

Eliza Meets Parry

Vincent Cerf, RFC439

lead

Good Evening. Tell me your problems. Please terminate input with a period or a question mark.

*People get on my nerves sometimes.

I am not sure I understand you fully.

*You should pay more attention.

Suppose you should pay more attention.

*You're entitled to your own opinion.

What makes you think I am entitled to my own opinion?

*It bothers me just to be around people in general.

Please go on.

*I went to the track at Bay Meadows a while back.

What does that suggest to you?

*I went to the races.

What else would you like to discuss?

*Horseracing interests me.

Why do you say that just now?

*There's a lot of crooked gambling going on there.

I am not sure I understand you fully,

*It's clear to me if not to you.

Do you think its likely that not to I?

*I went to the races.

Please go on.

*I gamble on horses.

What does that suggest to you?

*I went to the races.

What else would you like to discuss?

*Gambling has been nothing but trouble to me.

Why do you say that just now?

*I've had experience with crooks gambling at the track.

I am not sure I understand you fully.

*I've already told you enough about that.

Woebot

- Chatbot for delivering Cognitive Behavior Therapy via brief daily conversations
 - Little decision tree of language and menu responses

"What's going on in your world right now?"

"How are you feeling?"

In response to endorsed loneliness:

"I'm so sorry you're feeling lonely. I guess we all feel a little lonely sometimes"

- In a 2-week experiment seemed to reduce depression, but not anxiety

Fitzpatrick, Darcy, Vierhile, 2017. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health 4:2.

Simple Chatbot – Rule based Demo

- NLTK has a module, nltk.chat, which simplifies building Chatbots
- **Chat**: This is a **class** that has all the logic that is used by the chatbot.
- **Reflections**: This is a **dictionary** that contains a set of **input values and its corresponding output values**. It is an optional dictionary that you can use. You can also create your own dictionary in the same format as below and use it in your code.

IR-based chatbots

- Idea: Mine conversations of human chats or human-machine chats
 - Microblogs: Twitter or Weibo
 - Movie dialogs
- Cleverbot (Carpenter 2017
<http://www.cleverbot.com>)
- Microsoft Tay

A Cleverbot conversation

User:	Do you like Doctor Who?
Cleverbot:	Yes, so funny
User:	What's your favorite show?
Cleverbot:	I don't have one
User:	What's your name?
Cleverbot:	Jason I already told you.
User:	No your real name.
Cleverbot:	Stephanie IS my real name.
User:	I thought it was Jason?
Cleverbot:	How did you come to that conclusion.
User:	Because you said so.
Cleverbot:	I have to do my homework.

Two IR-based chatbot architectures

1. Return the response to the most similar turn

- Take user's turn (q) and find a (tf-idf) similar turn t in the corpus C
 $q = \text{"do you like Doctor Who"}$

$t' = \text{"do you like Doctor Strangelove"}$

- Grab whatever the response was to t .

$$r = \text{response} \left(\operatorname{argmax}_{t \in C} \frac{q^T t}{\|q\| \|t\|} \right)$$

Yes, so funny

2. Return the most similar turn

$$r = \operatorname{argmax}_{t \in C} \frac{q^T t}{\|q\| \|t\|}$$

Do you like Doctor Strangelove

IR-based models of chatbots

- Also fine to use other features like user features, or prior turns
- Or non-dialogue text
 - COBOT chatbot (Isbell et al., 2000)
 - sentences from the Unabomber Manifesto by Theodore Kaczynski, articles on alien abduction, the scripts of “The Big Lebowski” and “Planet of the Apes”.
 - Wikipedia text

IR based Chatbot Demo

- Web scraping using Python
- Retrieve Web page on Chatbots
- Bot answers questions related to the Web page content.

Neural chatbots

- Think of response generation as a task of *transducing* from the user's prior turn to the system's turn.
- Train on:
 - movie dialogue databases
 - Twitter conversations
- Train a deep neural network
 - map from user1 turn to user2 response

Sample neural chatbot output

Input	tell me ... how long have you had this falling sickness ?
Vanilla-SEQ2SEQ	i'm not a doctor.
Adversarial	a few months, i guess .
Input	so i had the doctors test sammy 's response to conditioning .
Vanilla-SEQ2SEQ	sammy wrote the test sammy wrote the test .
Adversarial	so he took the pills .
Input	they didn 't have much success with this problem commander .
Vanilla-SEQ2SEQ	they 're not the only ones who have been in the system .
Adversarial	can we find someone else ?

Adversarial neural network use generative models

Generate new data with the same statistics as the training set. For example, a GAN trained on photographs can generate new photographs that look at least superficially authentic

Oh I've never seen that! How long does it take you
? guys to learn the drill?
Like 2 weeks ago!!

Chatbots: pro and con

Pro:

- Fun
- Applications to counseling
- Good for narrow, scriptable applications

Cons:

- They don't really understand
- Rule-based chatbots are precise but expensive
- IR-based or neural networks chatbots can only mirror training data
 - (Garbage-in, Garbage-out)

The future: combining chatbots with frame-based agents

Frame-based dialog agents

- Sometimes called "task-based dialog agents"
- Based on a "domain ontology"
 - A knowledge structure representing user intentions
- One or more **frames**
 - Each a collection of **slots**
 - Each slot having a **value**

The Frame

- A set of **slots**, to be filled with information of a given **type**
- Each associated with a **question** to the user

Slot	Type	Question
ORIGIN	city	What city are you leaving from?
DEST	city	Where are you going?
DEP DATE	date	What day would you like to leave?
DEP TIME	time	What time would you like to leave?
AIRLINE	line	What is your preferred airline?

Slot types can be complex

The type *DATE*

DATE

MONTH NAME

DAY (BOUNDED-INTEGER 1 31)

YEAR INTEGER

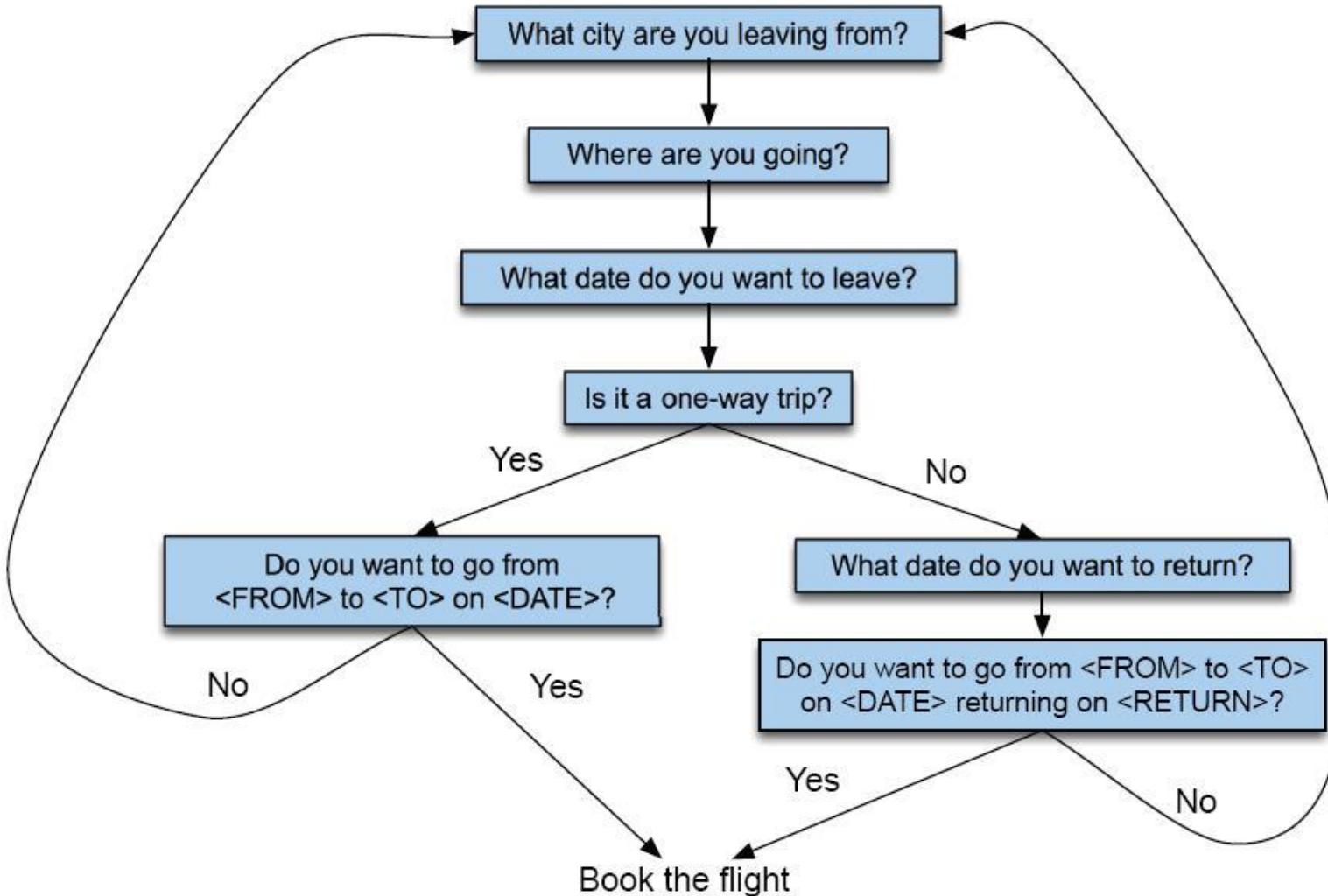
WEEKDAY (MEMBER (SUNDAY MONDAY TUESDAY WEDNESDAY THURSDAY FRIDAY SATURDAY)]

Control structure for frame-based dialog

Consider a trivial airline travel system:

- Ask the user for a departure city
- Ask for a destination city
- Ask for a time
- Ask whether the trip is round-trip or not

Finite State Dialog Manager



Finite-state dialog managers

- System completely controls the conversation with the user.
- It asks the user a series of questions
- Ignoring (or misinterpreting) anything the user says that is not a direct answer to the system's questions

Dialogue Initiative

- Systems that control conversation like this are called **single initiative**.
- **Initiative**: who has control of conversation
- In normal human-human dialogue, initiative shifts back and forth between participants.

System Initiative

System completely controls the conversation

- Simple to build
 - User always knows what they can say next
 - System always knows what user can say next
 - Known words: Better performance from ASR
 - Known topic: Better performance from NLU
 - OK for VERY simple tasks (entering a credit card, or login name and password)
-
- - Too limited



Problems with System Initiative

- Real dialogue involves give and take!
- In travel planning, users might want to say something that is not the direct answer to the question.
- For example answering more than one question in a sentence:

Hi, I'd like to fly from Seattle Tuesday morning

I want a flight from Milwaukee to Orlando one way leaving after 5 p.m. on Wednesday.

Single initiative + universals

- We can give users a little more flexibility by adding **universals**: commands you can say anywhere
 - As if we augmented every state of FSA with these
 - **Help**
 - **Start over**
 - **Correct**
 - This describes many implemented systems
 - But still doesn't allow user much flexibility
-

Instead, the GUS architecture

- GUS
 - Genial(friendly) Understannder System
 - A kind of ***mixed initiative***
 - The conversational initiative shifts between system and user
 - The structure of the **frame** guides dialogue

Frames are mixed-initiative

- System asks questions of user, filling any slots that user specifies
 - When frame is filled, do database query
- If user answers 3 questions at once, system can fill 3 slots and not ask these questions again!

Frame-based dialogue agents

1977:

versation with a client who wants to make a simple return trip to a single city in California.

There is good reason for restricting the domain of discourse for a computer system which is to engage in an English dialog. Specializing the subject matter that the system can talk about permits it to achieve some measure of realism without encompassing all the possibilities of human knowledge or of the English language. It also provides the user with specific motivation for participating in the conversation, thus narrowing the range of expectations that GUS must have about the user's purposes. A system restricted in this way will be more able to guide the conversation within the boundaries of its competence.

<https://nlp.stanford.edu/acvogel/gus.pdf>

Artificial Intelligence Journal, 1977

Still the industrial state of the art

- Many task-based bots use GUS architecture

The state of the art in 1977 !!!!

GUS: Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go ?

Client: I want to go to San Diego on May 28

GUS: What time do you want to leave?

Client: I must be in San Diego before 10 am

GUS: Would you like P.S.A. flight 102 that arrives at 9.15 am ?

Client: What is the next flight ?

GUS: Air California flight 310 that leaves at 8.30am and arrives at 10.00am

Client: I'll take the first one

GUS: In what name should I make a reservation, please?

Client: Dan Bobrow

GUS: I have confirmed the following flight: P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.

GUS: What date do you want to return on ?

Client: On Friday in the evening.

GUS: Would you like the flight that leaves at 7.45 pm ?

Client: That's fine.

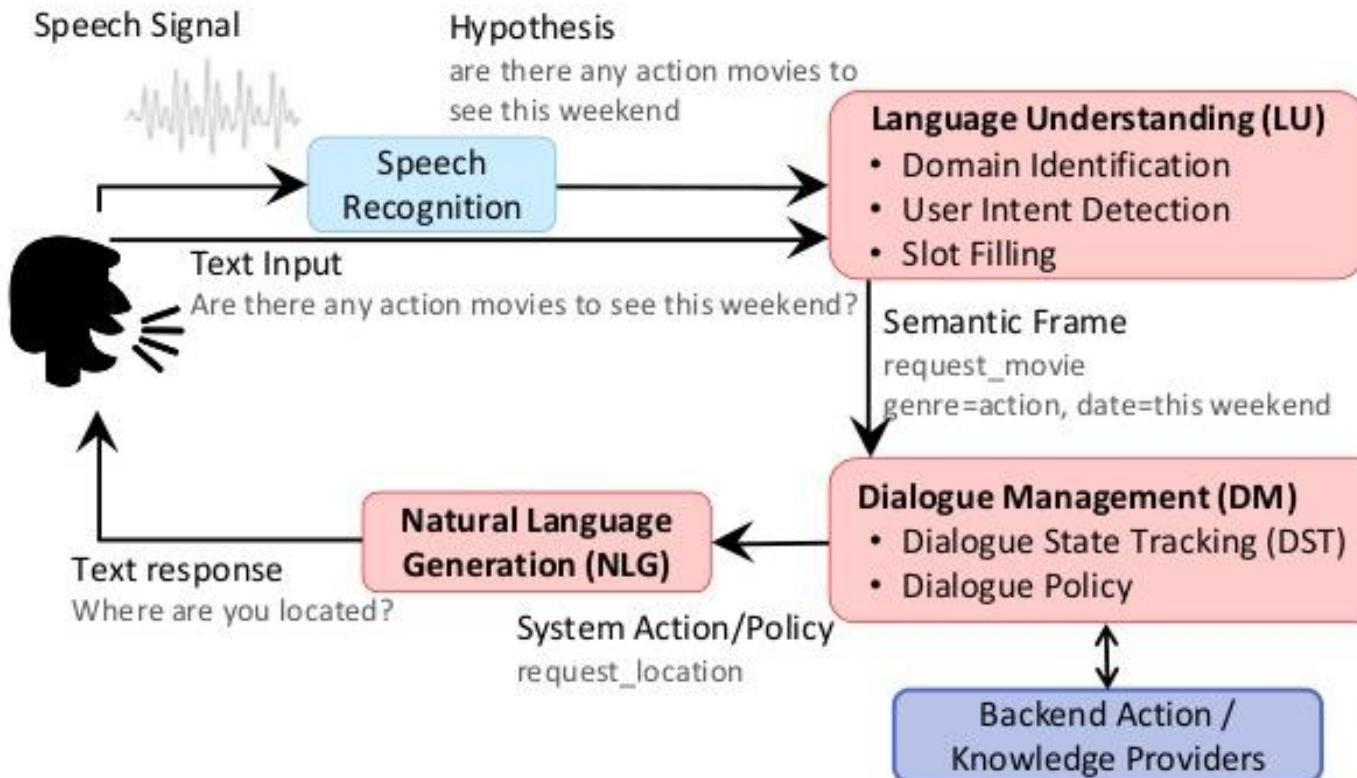
GUS: I have confirmed the following flight: P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm Thank you for calling. Goodbye

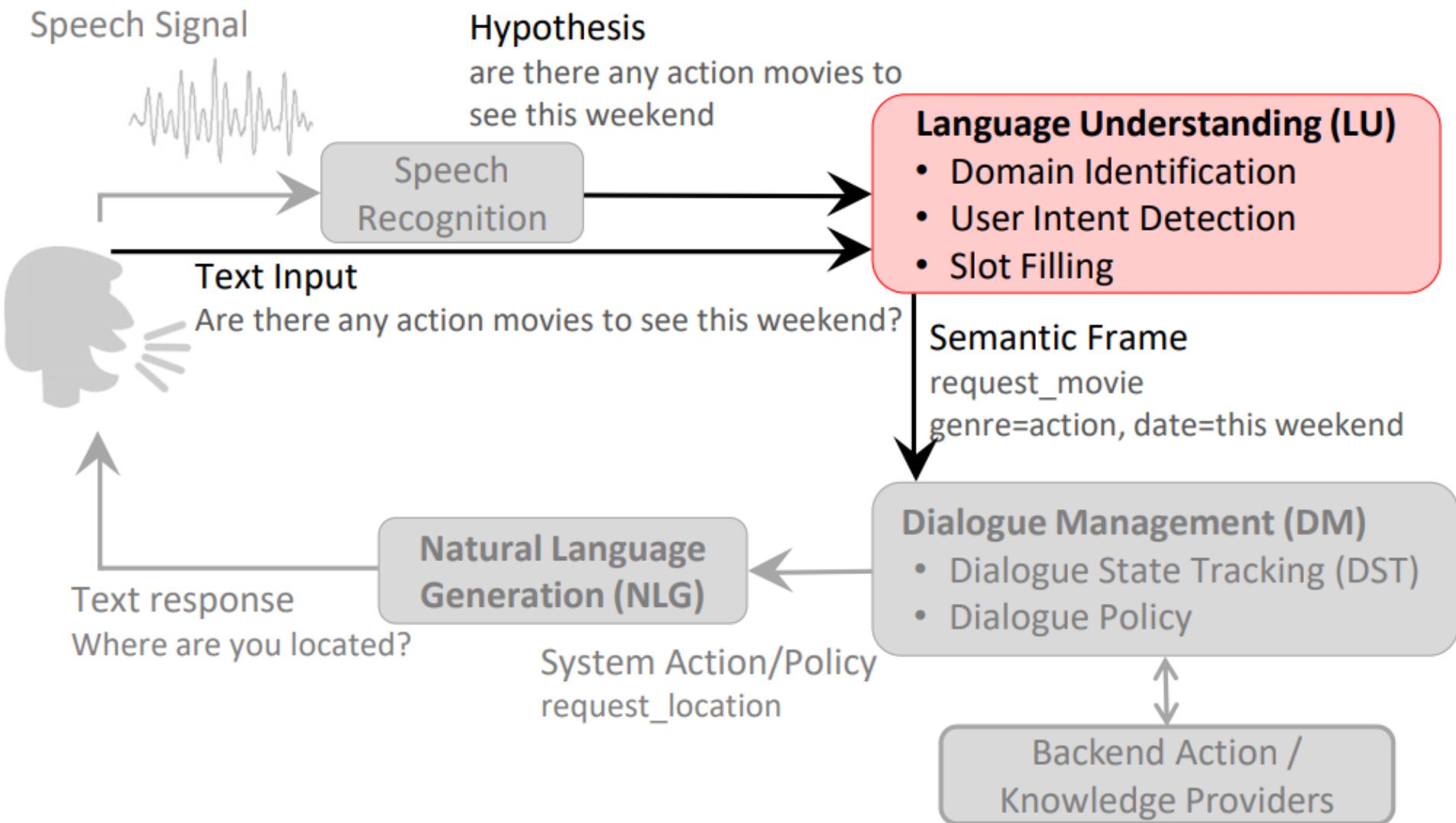
Task oriented dialogue agents

Task-Oriented Dialogue System (Young, 2000)

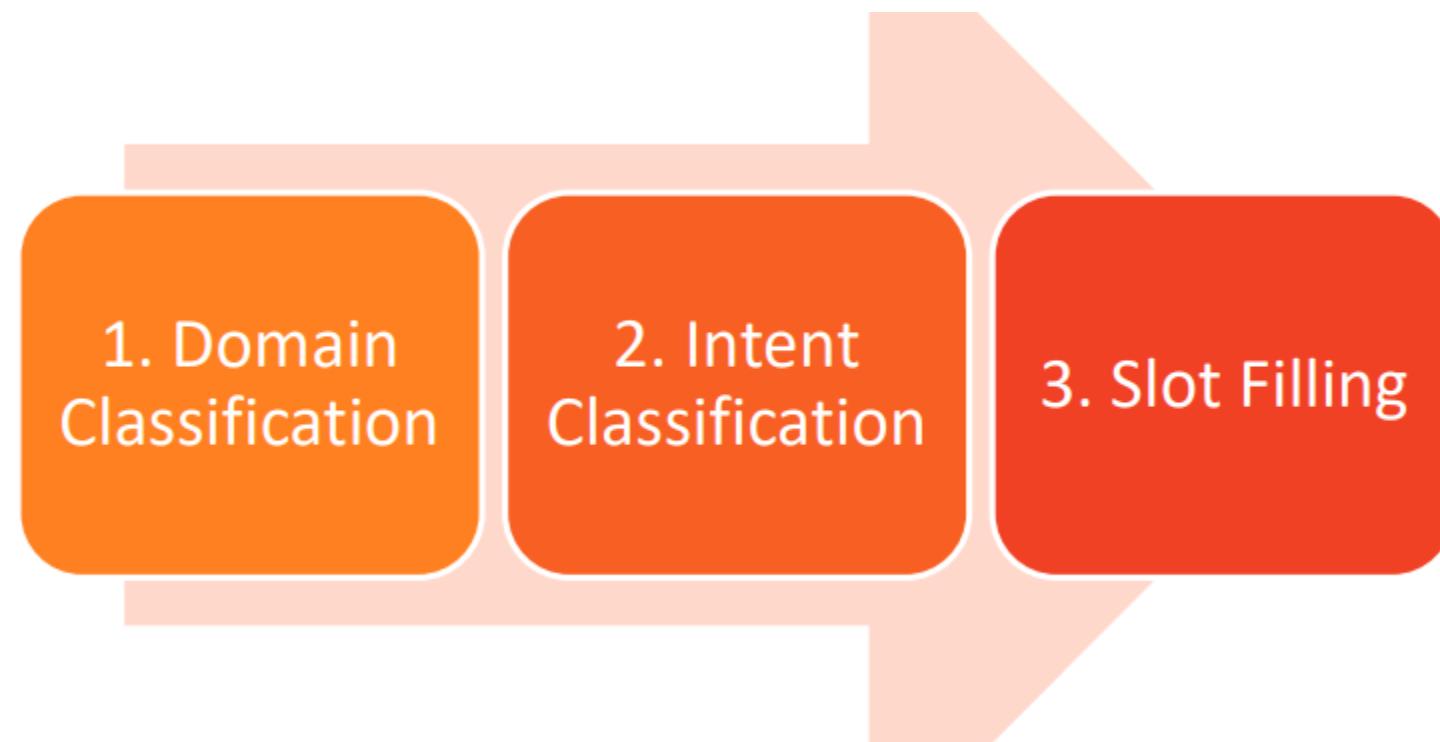
12

<http://rsta.royalsocietypublishing.org/content/358/1769/1389.short>





Language Understanding



Natural Language Understanding for filling dialog slots



1. Domain classification

Asking weather? Booking a flight?
Programming alarm clock?

2. Intent Determination

Find a Movie, Show Flight, Remove
Calendar Appt

3. Slot Filling

Extract the actual slots and fillers

Natural Language Understanding for filling slots

Show me morning flights from Boston to SF on Tuesday.

DOMAIN:	AIR-TRAVEL
INTENT:	SHOW-FLIGHTS
ORIGIN-CITY:	Boston
ORIGIN-DATE:	Tuesday
ORIGIN-TIME:	morning
DEST-CITY:	San Francisco

Natural Language Understanding for filling slots

Wake me tomorrow at six.

DOMAIN: ALARM-CLOCK

INTENT: SET-ALARM

TIME: 2017-07-01 0600-0800

Rule-based Slot-filling

Write regular expressions or grammar rules

Wake me (up) | set (the|an) alarm | get me up

Do text normalization

Siri uses GUS architecture: Condition-Action Rules

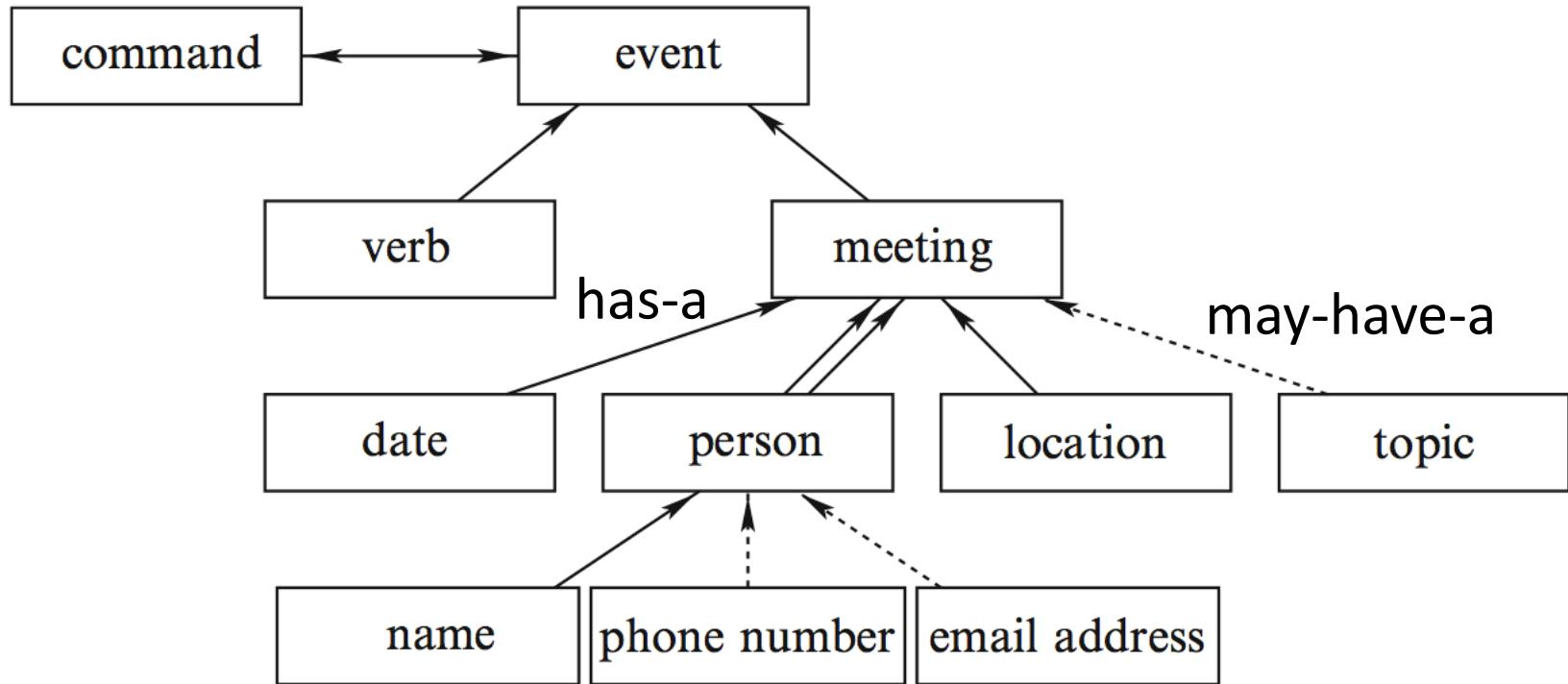
Active Ontology: relational network of concepts

- **data structures**: a **meeting** has
 - a date and time,
 - a location,
 - a topic
 - a list of attendees
- **rule sets** that perform actions for concepts
 - the **date** concept turns string
 - *Monday at 2pm* into
 - date object

Rule sets

- Collections of **rules** consisting of:
 - condition
 - action
- When user input is processed, facts added to store and
 - rule conditions are evaluated
 - relevant actions executed

Part of ontology for meeting task



meeting concept: if you don't yet have a location, ask for a location

Machine learning for slot-filling:

- Machine learning classifiers to map words to semantic frame-fillers
- Given a set of labeled sentences
 - “I want to fly to San Francisco on Tuesday”
 - Destination: SF
 - Depart-date: Tuesday
 - Build a classifier to map words to the slots
- Requirements: Lots of labeled data



Machine learning for slot-filling: Domain and Intent

I want to fly to San Francisco on Monday afternoon please

Use 1-of-N classifier (naive bayes, logistic regression, neural network, etc.)

- Input:
features like word N-grams

- Output:
Domain: AIRLINE
Intent: SHOWFLIGHT

Machine learning for slot-filling: Slot presence



I want to fly to San Francisco on Monday afternoon please

Use 1-of-N classifier (naive bayes, logistic regression, neural network, etc.)

- Input:
features like word N-grams, gazetteers (lists of cities)
- Output:
Destination-City

Machine learning for slot-filling: Slot filler



I want to fly to San Francisco on Monday afternoon please

Use 1-of-N classifier (naive bayes, logistic regression, neural network, etc.) for Destination City

- Input:
 - features like word N-grams, gazetteers (lists of cities)
- Output:
 - San Francisco

More sophisticated algorithm for slot filling: IOB/ BIO Tagging

- IOB or BIO Tagging
 - tag for the beginning (B) and inside (I) of each slot label,
 - plus one for tokens outside (O) any slot label.
 - $2n + 1$ tags, where n is the number of slots.

B-DESTINASTION

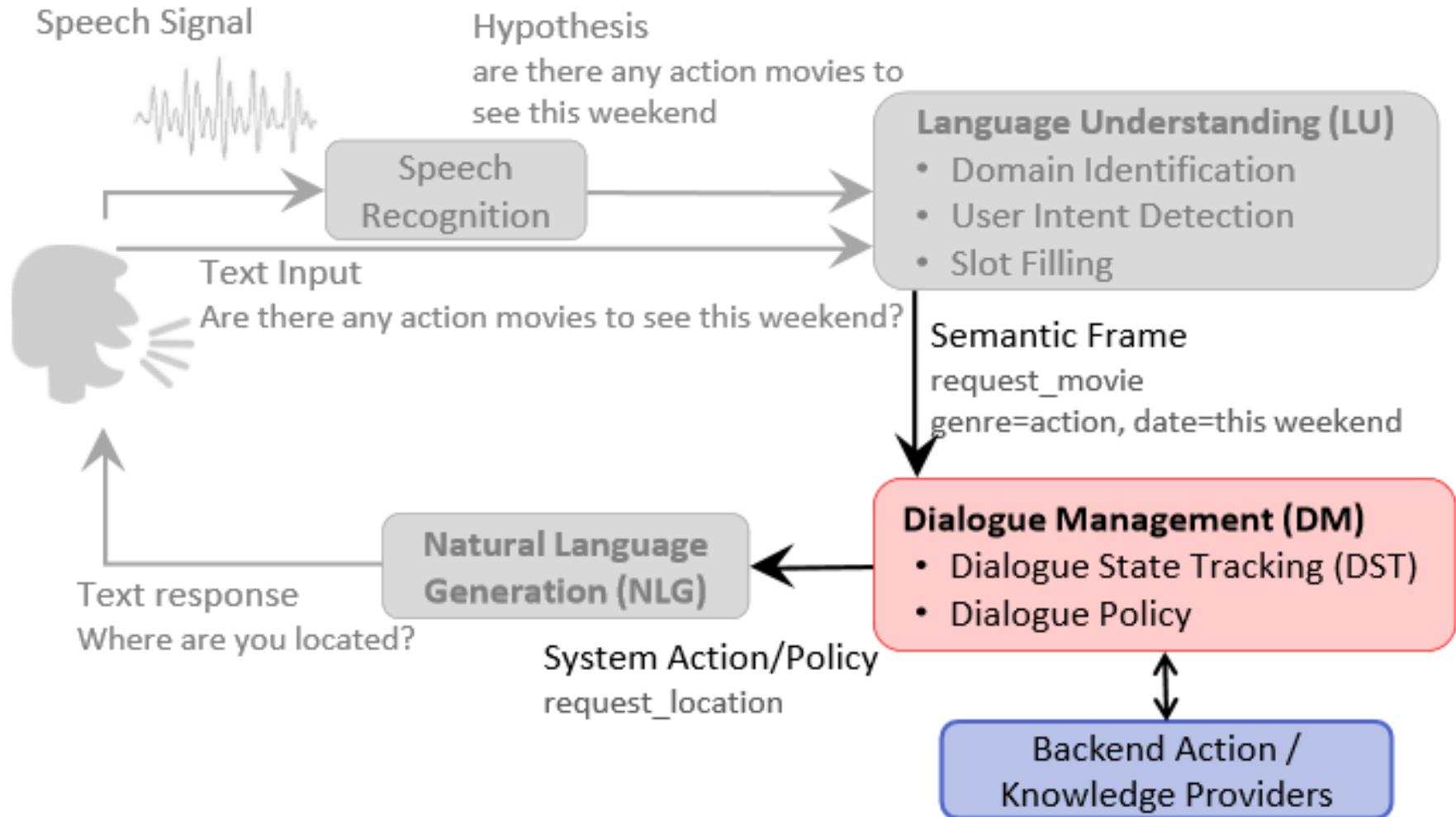
I-DESTINATION

B-DEPART_TIME

I-DEPART_TIME

0 0 0 0 B-DES I-DES 0 B-DEPTIME I-DEPTIME 0
I want to fly to San Francisco on Monday afternoon please

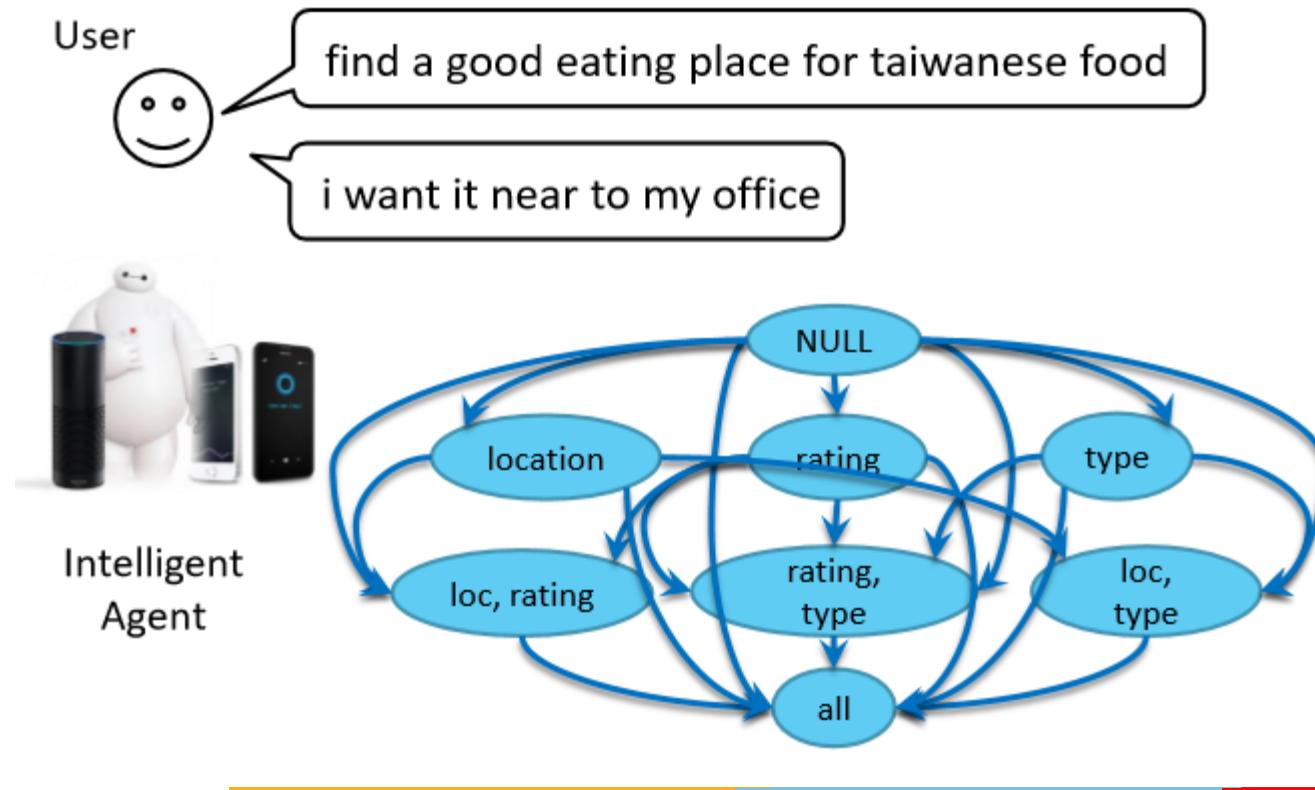
Dialog Management



Dialog Manager : Dialog State tracking

State Tracking

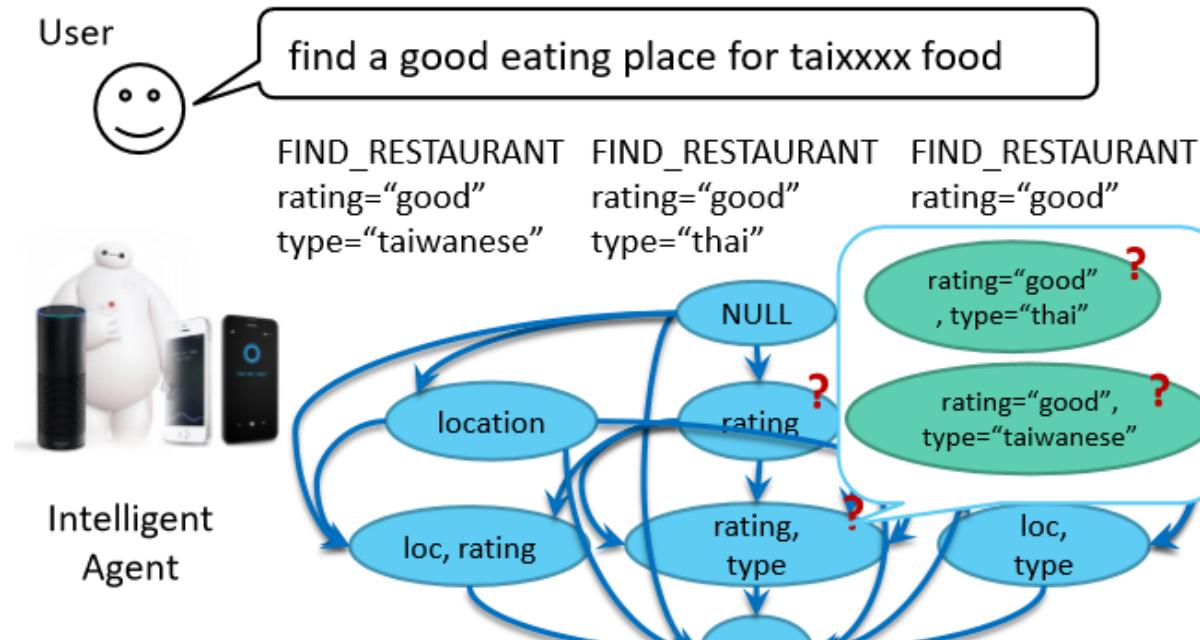
Requires Hand-Crafted States



Dialog Manager : Dialog State tracking

State Tracking

Handling Errors and Confidence



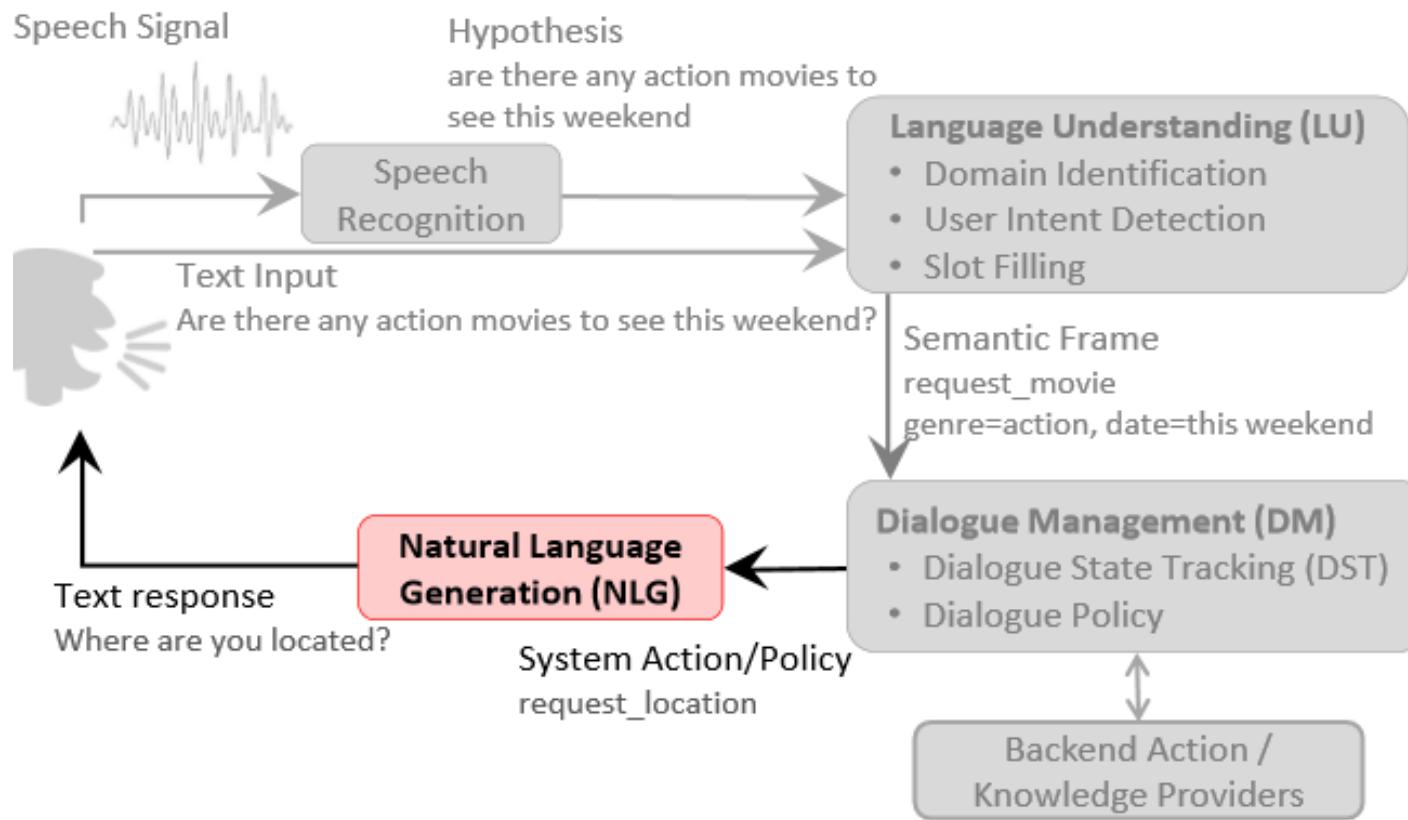
Dialog Manager : Dialog Policy

Dialogue Policy for Agent Action

- Inform(location="Taipei 101")
 - ▣ "The nearest one is at Taipei 101"
- Request(location)
 - ▣ "Where is your home?"
- Confirm(type="taiwanese")
 - ▣ "Did you want Taiwanese food?"

Natural Language Generation

Task-Oriented Dialogue System (Young, 2000)



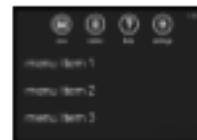
Natural Language Generation

Output / Natural Language Generation

- Goal: generate natural language or GUI given the selected dialogue action for interactions



- Inform(location="Taipei 101")
 - "The nearest one is at Taipei 101" v.s.
- Request(location)
 - "Where is your home?" v.s.
- Confirm(type="taiwanese")
 - "Did you want Taiwanese food?" v.s.



Other components of SIRI-style architectures

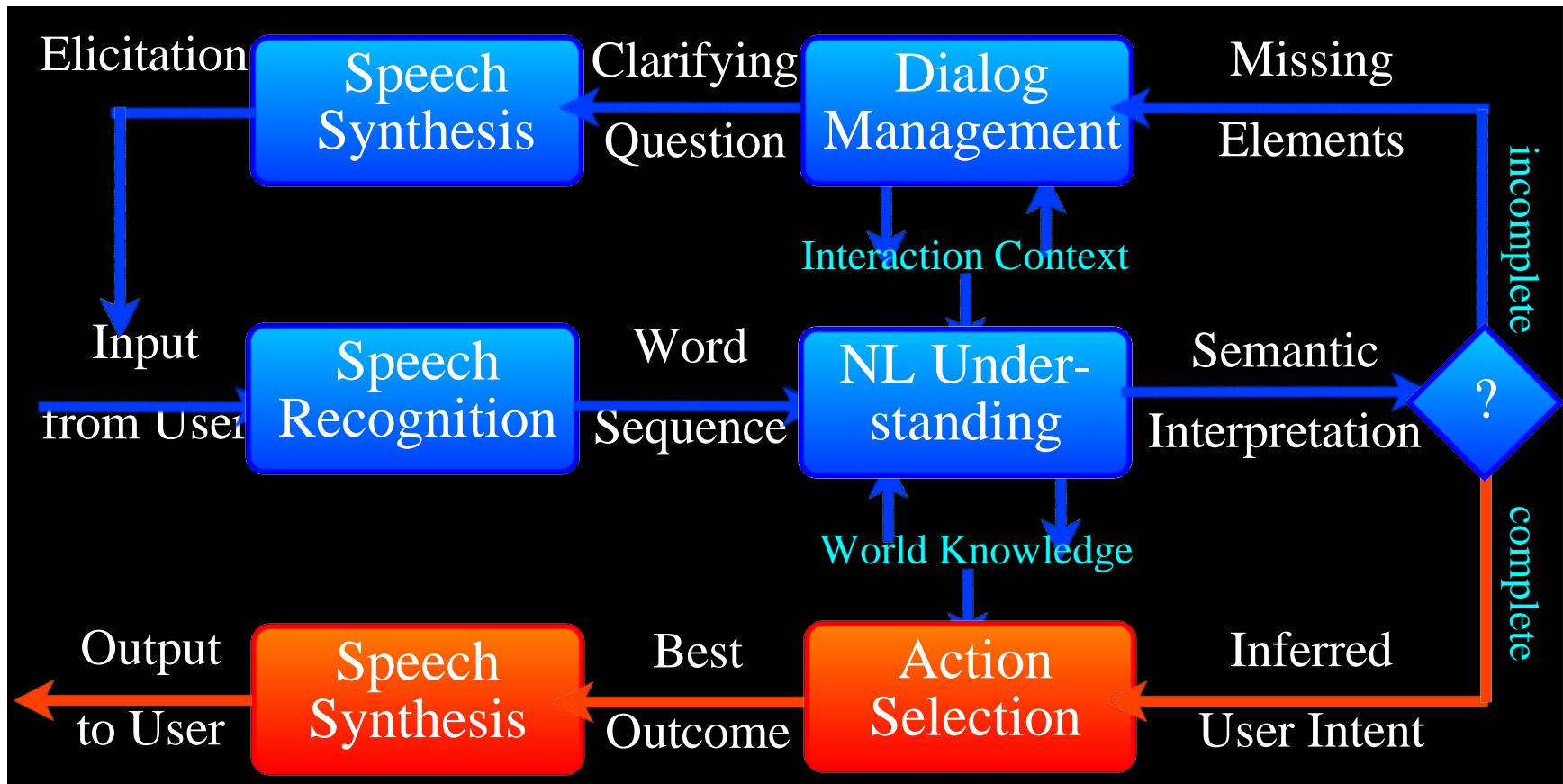
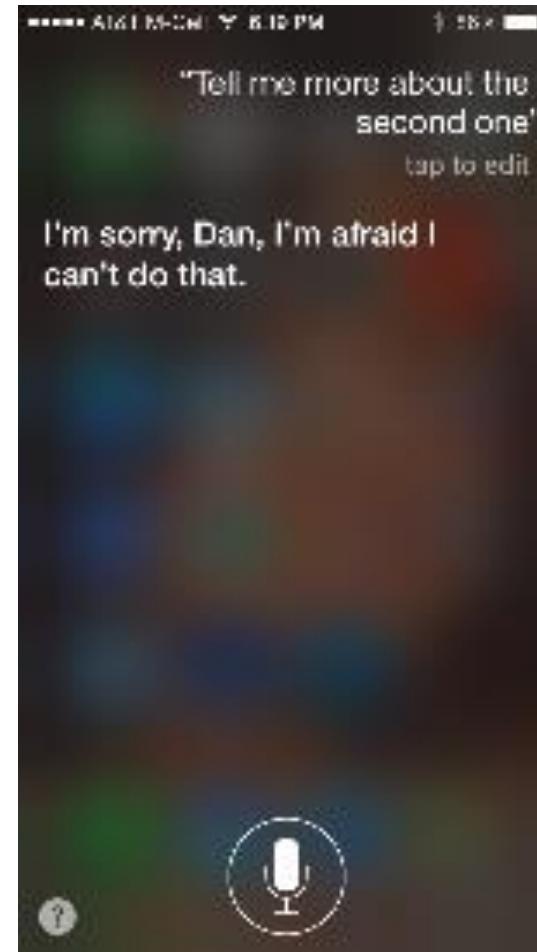
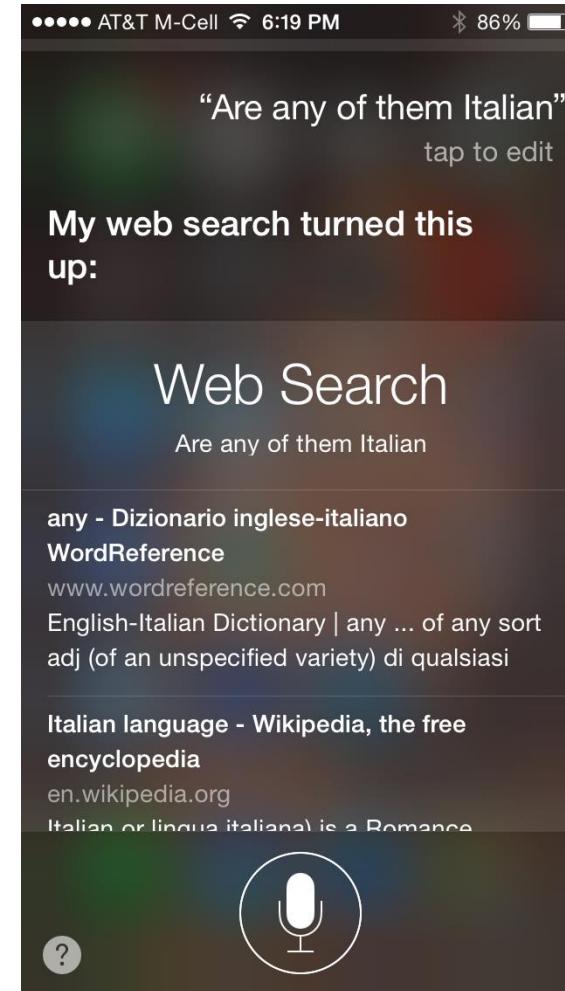


Figure from Jerome Bellegarda

SIRI 2014





SIRI 2017



••••• AT&T M-Cell ⌂ 10:52 AM 100% 🔋

"Find restaurants near me"
tap to edit

Here's what I found:

15 Results
Nearby

Ichи Sushi & Ni Bar
Japanese · 800 feet · Closed now
★★★★★ (311) on Yelp · \$\$\$

Coco's Ramen
Vietnamese · 800 feet · Closed now
★★★★★ (280) on Yelp · \$\$

Blue Plate
Restaurant · 0.2 miles · Closed now
★★★★★ (1251) on OpenTable · \$\$\$

Jasmine Tea House
Chinese · 900 feet · Closed now

?

Microphone icon

••••• AT&T M-Cell ⌂ 10:53 AM 100% 🔋

"Tell me more about the first one"
tap to edit

Ichи Sushi & Ni Bar is a Japanese restaurant on Mission St 800 ft to your west. It averages 3½ stars and is somewhat pricey.

SIRI 2017



AT&T M-Cell 10:52 AM 100%

"Find restaurants near me"
tap to edit

Here's what I found:

15 Results
Nearby

Ichisushi & Ni Bar
Japanese · 800 feet · Closed now
★★★★★ (311) on Yelp · \$\$\$

Coco's Ramen
Vietnamese · 800 feet · Closed now
★★★★★ (280) on Yelp · \$\$

Blue Plate
Restaurant · 0.2 miles · Closed now
★★★★★ (1251) on OpenTable · \$\$\$

Jasmine Tea House
Chinese · 900 feet · Closed now

?

?

AT&T M-Cell 10:54 AM 100%

"Are any of them Italian"
tap to edit

OK, here's what I found:

15 Results
Nearby

Emmy's Spaghetti Shack
Italian · 0.2 miles · Closed now
★★★★★ (233) on Yelp · \$\$

Vega
Pizza · 0.2 miles · Closed now
★★★★★ (423) on Yelp · \$\$

Pizza Hut
Pizza · 800 feet
★★★★★ (69) on Yelp · \$

La Ciccia
Sardinian · 0.4 miles

?

?

QA Bot – NLP DeepLearning

- We will be implementing a chat bot that can answer questions based on a given story.
- **The bAbI project** by Facebook research
- <https://research.fb.com/downloads/babi/>

This page gather resources related to the bAbI project of Facebook AI Research which is organized towards the goal of automatic text understanding and reasoning.

QA Bot

- Story
 - Jane went to the store. Mike ran to the bedroom.
- Question
 - Is Mike in the store?
- Answer
 - No

QA Bot

End-to-End Memory Networks

- Sainbayar Sukhbaatar
- Arthur Szlam
- Jason Weston
- Rob Fergus

**You must read the paper to understand
this network!**

[1503.08895.pdf \(arxiv.org\)](https://arxiv.org/pdf/1503.08895.pdf)

QA Bot

- Model takes a discrete set of inputs $\mathbf{x}_1, \dots, \mathbf{x}_n$ that are to be stored in the memory, a query \mathbf{q} , and outputs an answer \mathbf{a}
- Each of the \mathbf{x} , \mathbf{q} , and \mathbf{a} contains symbols coming from a dictionary with V words.
- The model writes all \mathbf{x} to the memory up to a fixed buffer size, and then finds a continuous representation for the \mathbf{x} and \mathbf{q} .

QA Bot- NN Demo

- Training data is a list of tuples.
- Each tuple has a Story, Question and an Answer.
- Training and Testing Data should be a part of Vocabulary.

- Padding: Pad_sequences

```
sample_text_1="bitty bought a bit of butter"
sample_text_2="but the bit of butter was a bit bitter"
sample_text_3="so she bought some better butter to make the bitter butter bette
r"
```

```
The encoding for document 1 is : [45, 16, 32, 27, 34, 33]
The encoding for document 2 is : [24, 2, 27, 34, 33, 37, 32, 27, 3]
The encoding for document 3 is : [22, 27, 16, 28, 35, 33, 7, 2, 2, 3, 33, 3
5]

The padded encoding for document 1 is : [45 16 32 27 34 33 0 0 0 0 0
0]
The padded encoding for document 2 is : [24 2 27 34 33 37 32 27 3 0 0
0]
The padded encoding for document 3 is : [22 27 16 28 35 33 7 2 2 3 33 3
5]
```

Evaluation

- Natural Language Understanding component Evaluation:
 1. Slot Error Rate for a Sentence =
$$\frac{\# \text{ of inserted/deleted/substituted slots}}{\# \text{ of total reference slots for sentence}}$$
Slot tagging is most important since it contributes maximum to quality of Dialog System
 2. Intent Accuracy: Correct intent identified
- Dialog Manager Component Evaluation
 1. State tracking Accuracy
 2. End-to-end evaluation (Task Success)

Evaluation Metrics

“Make an appointment with Chris at 10:30 in Gates 104”

Slot	Filler
PERSON	Chris
TIME	11:30 a.m.
ROOM	Gates 104

Slot error rate: 1/3

Task success: At end, was the correct meeting added to the calendar?



Dialog System Design: User-centered Design

1. Study the user and task
2. Iteratively test the design on users

Ethical Issues in Dialog System Design

- Machine learning systems replicate biases that occurred in the training data.
- Microsoft's Tay chatbot
 - Went live on Twitter in 2016
 - Taken offline 16 hours later
- In that time it had started posting racial slurs, conspiracy theories, and personal attacks
 - Learned from user interactions (Neff and Nagy 2016)

Ethical Issues in Dialog System Design

- Machine learning systems replicate biases that occurred in the training data.
- Dialog datasets
 - Henderson et al. (2017) examined standard datasets (Twitter, Reddit, movie dialogs)
 - Found examples of hate speech, offensive language, and bias
 - Both in the original training data, and in the output of chatbots trained on the data.
- Fairness in Machine Learning

Ethical Issues in Dialog System Design: Privacy



- Remember this was noticed in the days of Weizenbaum
- Agents may record sensitive data
 - (e.g. “Computer, turn on the lights [answers the phone –Hi, yes, my password is...”],
- Which may then be used to train a seq2seq conversational model.
- Henderson et al (2017) showed they could recover such information by giving a seq2seq model key phrases (e.g., "password is")

Ethical Issues in Dialog System

Design: Gender equality

- Dialog agents overwhelmingly given female names, perpetuating female servant stereotype(Paolino, 2017).



Summary

Task-Oriented Bot

- Personal assistant, helps users achieve a certain task
- Combination of rules and statistical components
 - ▣ POMDP for spoken dialog systems (Williams and Young, 2007)
 - ▣ End-to-end trainable task-oriented dialogue system (Wen et al., 2016)
 - ▣ End-to-end reinforcement learning dialogue system (Li et al., 2017; Zhao and Eskenazi, 2016)



Chit-Chat Bot

- No specific goal, focus on natural responses
- Using variants of seq2seq model
 - ▣ A neural conversation model (Vinyals and Le, 2015)
 - ▣ Reinforcement learning for dialogue generation (Li et al., 2016)
 - ▣ Conversational contextual cues for response ranking (AI-Rfou et al., 2016)



Summary

- Chatbots
 - Simple rule-based systems
 - IR or Neural networks: mine datasets of conversations.
- Task-based Dialog System
 - Natural Language Understanding
 - Domain Classification
 - Intent Classification
 - Slot tagging
 - Dialog Manager
 - State Tracking
 - Dialog Policy
- Frame-based systems:
 - Hand-written rules for slot fillers
 - ML classifiers or Neural networks to fill slots

Summary



Human-Robot interfaces is a hot topic but several components must be integrated!

Most state-of-the-art technologies are based on DNN

- Requires huge amounts of labeled data
- Several frameworks/models are available



Fast domain adaptation with scarce data + re-use of rules/knowledge



Handling reasoning



Data collection and analysis from un-structured data



Complex-cascade systems requires high accuracy for working good as a whole

<https://www.topbots.com/building-conversational-ai/>

https://www.mindmeld.com/docs/intro/approaches_for_building_conversational_applications.html

References

- Speech and Language processing: An introduction to Natural Language Processing, Computational Linguistics and speech Recognition by Daniel Jurafsky and James H. Martin[3rd edition] - Chapter 26 Page 495
- <https://www.sciencedirect.com/science/article/abs/pii/0004370277900182>
- <https://blog.kore.ai/conversational-ai-top-20-trends-for-2020>
- <https://chatbotsmagazine.com/tutorials/home>
- <https://rasa.com/docs/rasa/user-guide/building-assistants/>

Training Data

- <https://rasa.com/docs/rasa/nlu/training-data-format/>
- <https://rasa.com/docs/rasa/user-guide/command-line-interface/#train-a-model>

NLU Components: <https://rasa.com/docs/rasa/core/slots/>

- <https://rasa.com/docs/rasa/nlu/components/>

Supporting Multiple languages: <https://rasa.com/docs/rasa/nlu/language-support/>

Voice Bots : <https://rasa.com/docs/rasa/user-guide/messaging-and-voice-channels/>

- <https://blog.rasa.com/connect-your-rasa-ai-assistant-to-amazon-alexa/>
- <https://blog.rasa.com/integrating-rasa-with-knowledge-bases/>



Natural Language Processing Applications



BITS Pilani
Pilani Campus

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 6: Knowledge Graph Use

Date – 28th December 2023

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Philip Cimiano , Johanna Völker , Paul Buitelaar and many others who made their course materials freely available online.

Session Content

What is a Knowledge Graph?

How to build a Knowledge Graph

How to Use Knowledge Graphs

- Merging Artificial Intelligence and Internet
- Knowledge Access Layer
- Open and Service-oriented Dialog Systems

Why we need Knowledge Graphs

- Motivation and Solution
- Touristic Use Cases
- Energy Use Cases
- Further Verticals

The Semantic Web Vision

“... the idea of having data on the Web defined and linked in a way that it can be used by machines not just for **display purposes**, but for **automation**, **integration** and **reuse** of data across various applications”



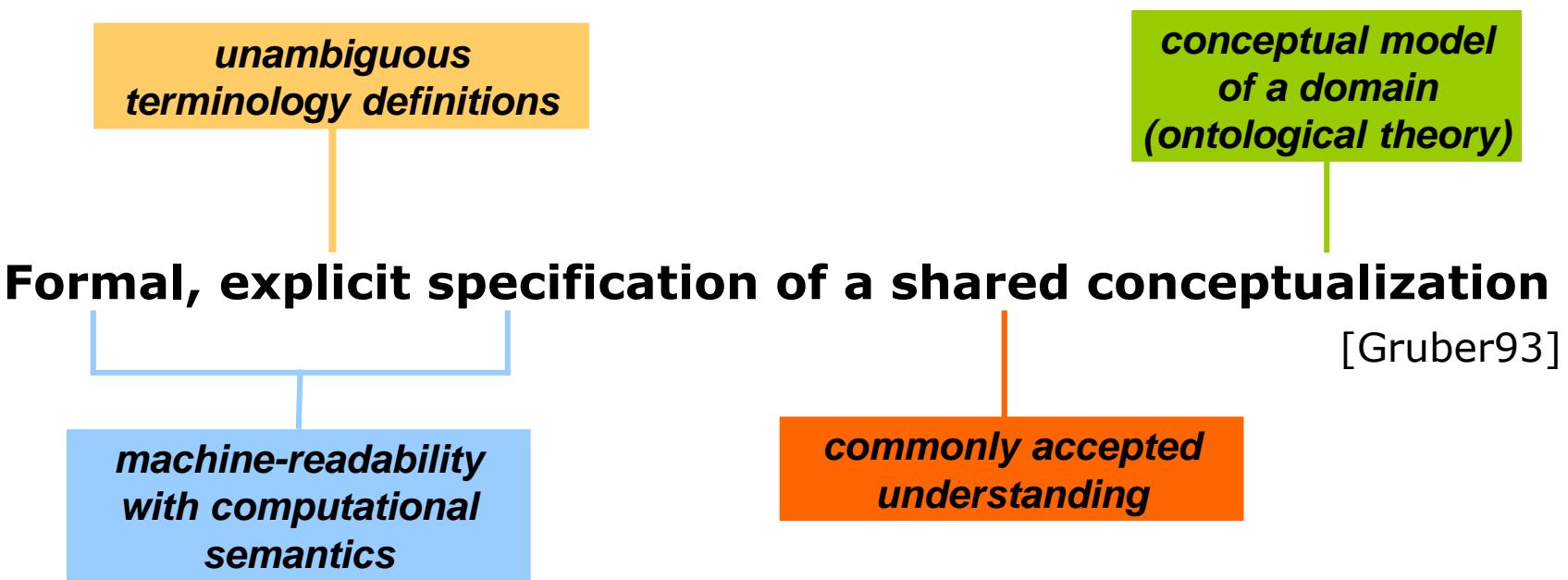
<http://www.w3.org/sw/>



Need to Add “Semantics”

- Use **Ontologies** to specify meaning of annotations
 - Ontologies provide a vocabulary of terms
 - New terms can be formed by combining existing ones
 - Meaning (**semantics**) of such terms is formally specified
 - Can also specify relationships between terms in multiple ontologies

Ontology Definition



Ontology Example

Concept

conceptual entity of the domain

Attribute

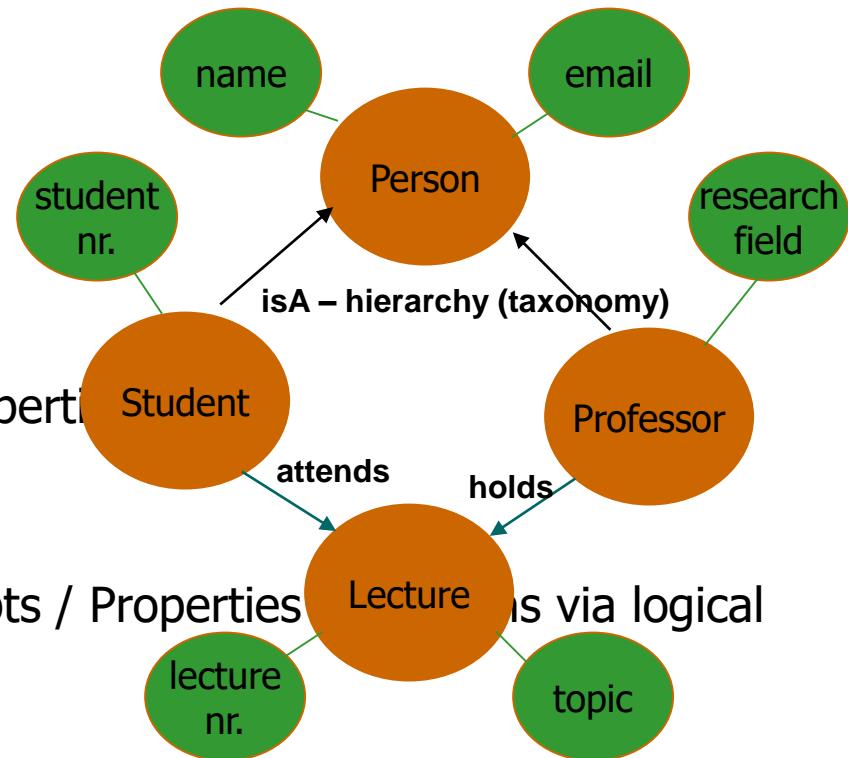
property of a concept

Relation

relationship between concepts or properties

Axiom

coherent description between Concepts / Properties
expressions

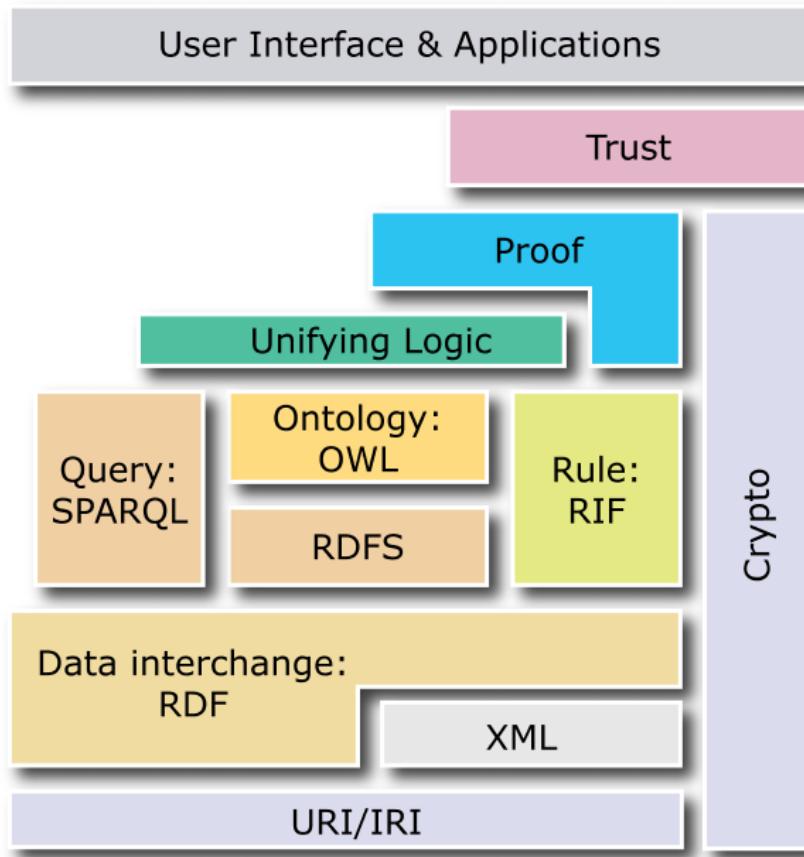


$\text{holds}(\text{Professor}, \text{Lecture}) \Rightarrow \text{Lecture}.\text{topic} \in \text{Professor}.\text{researchField}$

Ontology Languages

- For the purpose of formalizing ontologies, variants of first-order logic with standard model-theoretic semantics are used
 - RDF (Resource Description Framework)
 - Specifies relationship between data
 - RDFS(Resource Description Framework Schema)
 - Specifies relationship between schema
 - OWL (Web Ontology Language)
 - Specifies more complex relationship between schema based on description logics

Semantic Web Layers



URI/IRI

Universal Resource Identifier
Internationalized Resource Identifier

XML

eXtended Markup Language

RDF

Resource Description Framework

RDFS

RDF Schema

OWL

Web Ontology Language

SPARQL

Simple Protocol and RDF Query Language

Ontology Languages Summary

- RDF is a flexible data model for Semantic Web
- RDF Schema provides simple inference capability
- OWL allows more expressive representation of knowledge but is hard to scale to Web data
- Semantic technologies have been adopted by major companies such as Google, Yahoo and Facebook

Knowledge Graph

- **ontology + data = knowledge graph**
- Information extraction techniques for performing [named entity recognition](#). Another subfield which has gained much interest from the community is [keywords extraction](#).
- A **knowledge graph** is a way of storing data that resulted from an information extraction task. Many basic implementations of knowledge graphs make use of a concept we call **triple**, that is a set of three items(a subject, a predicate and an object)

Knowledge is the key

- **Semantic descriptions** of entities and their **relationships**
 - Uses a **knowledge representation formalism**
(Focus here: RDF, RDF-Schema, OWL)
 - **Entities**: real world objects (things, places, people) and abstract concepts (genres, religions, professions)
 - **Relationships**: graph-based data model where relationships are first-class
 - **Semantic descriptions**: types and properties with a well-defined meaning (e.g. through an ontology)
 - Possibly axiomatic knowledge (e.g. rules) to support automated reasoning
-

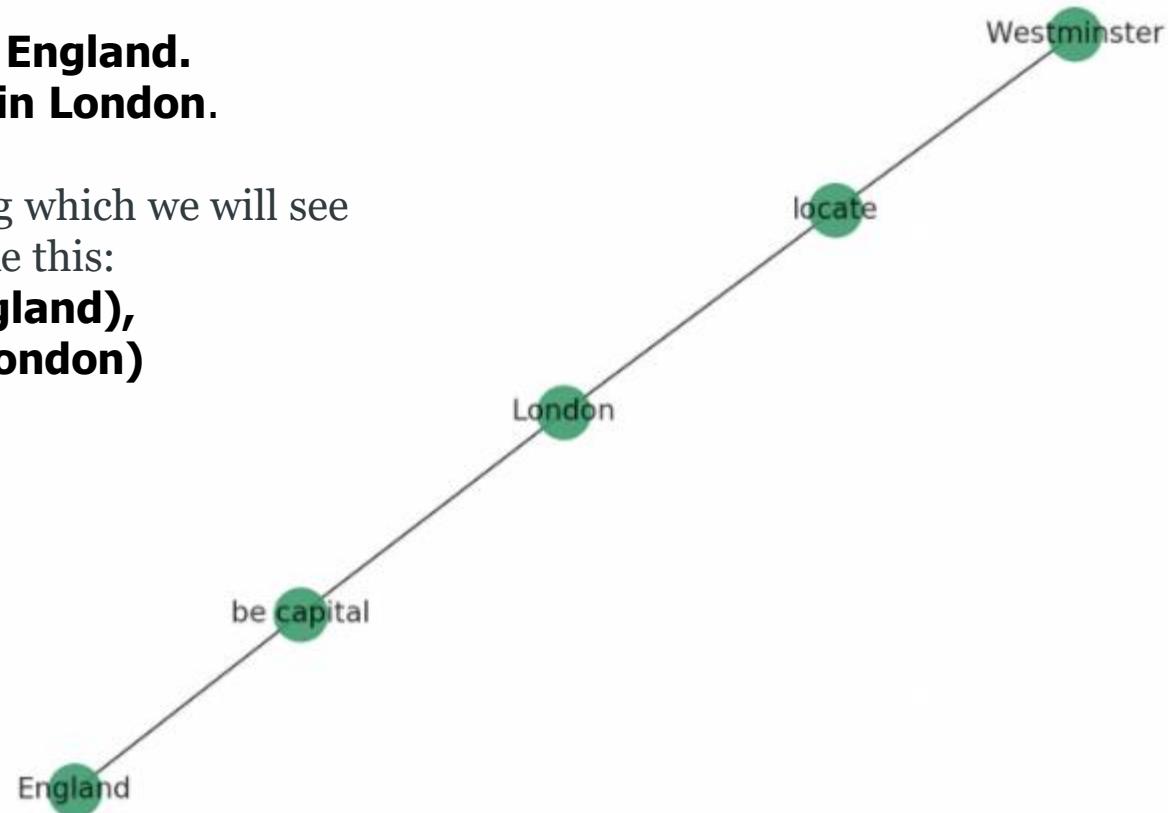
Example

London is the capital of England.

Westminster is located in London.

After some basic processing which we will see later, we would 2 triples like this:

**(London, be capital, England),
(Westminster, locate, London)**



Knowledge graph representation

Knowledge Graph technologies

Knowledge Graph technologies consists of:

- knowledge representation and reasoning (languages, schema and standard vocabularies),
- knowledge storage (graph databases and repositories),
- knowledge engineering (methodologies, editors and design patterns), and
- knowledge learning, including schema learning and population

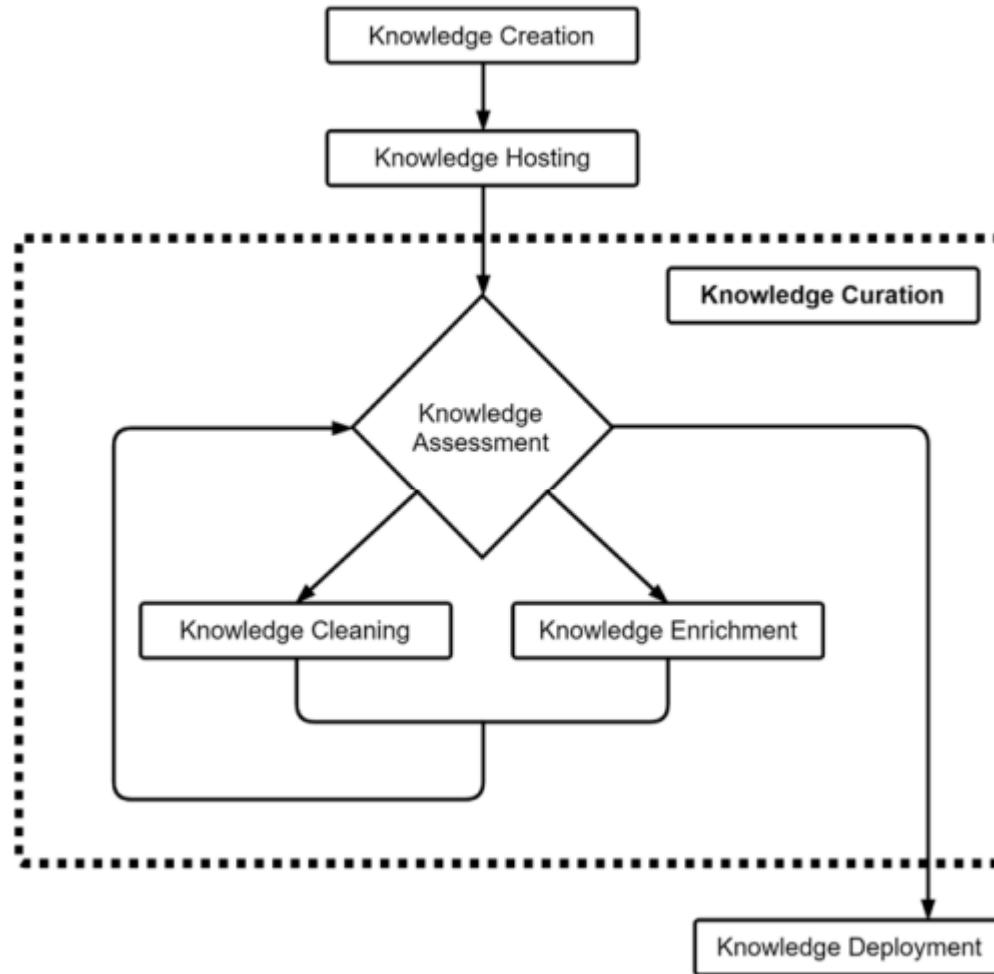
Key principles of Knowledge Graph



Building Knowledge Graph

- Identifying the entities and the relation between them is not a difficult task for us. However, **manually building a knowledge graph is not scalable.**
- Nobody is going to go through thousands of documents and extract all the entities and the relations between them!
- **To build a knowledge graph from the text, it is important to make our machine understand natural language.**
- This can be done by using NLP techniques such as sentence segmentation, dependency parsing, parts of speech tagging, and entity recognition

Knowledge Graph Generation Process



Knowledge Graph Generation

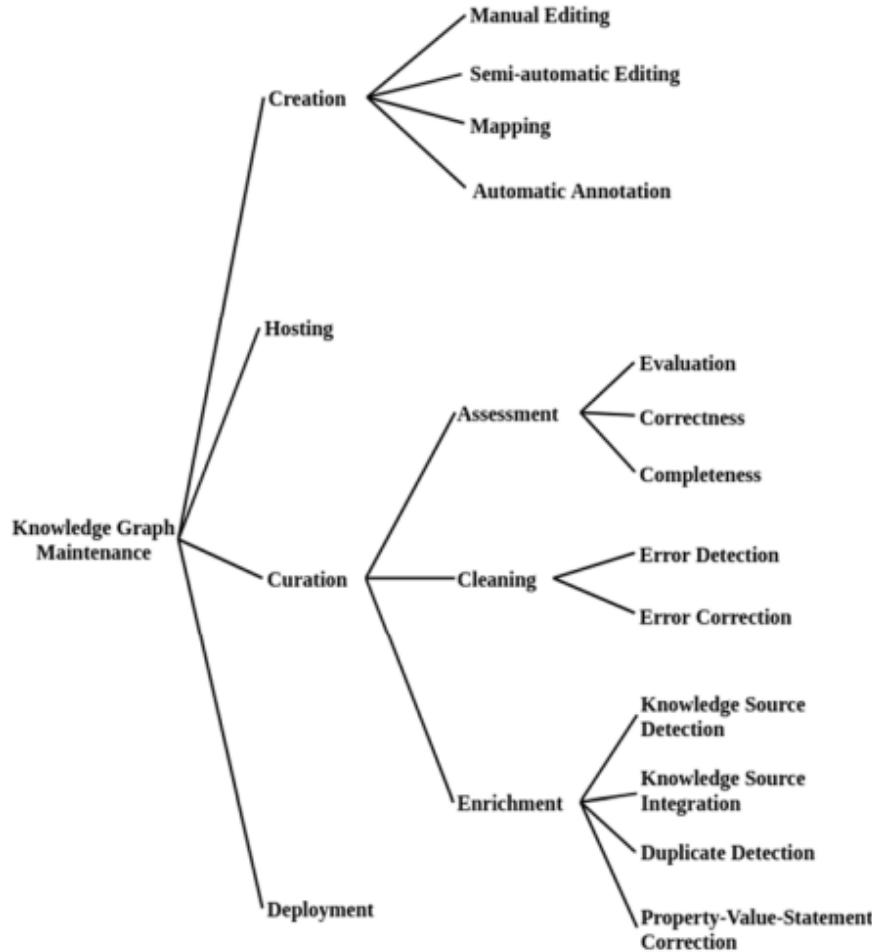


Fig. 2.2. A task model for Knowledge Graph Generation.

Linked Open Data Project

- Goal: “expose” open datasets in RDF
- Set RDF links among the data items from different datasets
- Set up, if possible, query endpoints

Example data source: DBpedia

- DBpedia is a community effort to
 - extract structured (“infobox”) information from Wikipedia
 - provide a query endpoint to the dataset
 - interlink the DBpedia dataset with other datasets on the Web



UNIVERSITÄT LEIPZIG



Extracting structured data from Wikipedia

```
@prefix dbpedia <http://dbpedia.org/resource/>.  
@prefix dbterm <http://dbpedia.org/property/>.
```

dbpedia:Amsterdam

```
dbterm:officialName "Amsterdam" ;  
dbterm:longd "4" ;  
dbterm:longm "53" ;  
dbterm:longs "32" ;  
dbterm:leaderName dbpedia:Lodewijk_Asscher ;  
...  
dbterm:areaTotalKm "219" ;
```

dbpedia:ABN_AMRO

```
dbterm:location dbpedia:Amsterdam ;  
...  
...
```

Amsterdam	
	
The Keizersgracht at dusk	
Location of Amsterdam	
Coordinates:	52°22'23"N 4°53'32"E
Country	Netherlands
Province	North Holland
Government	
- Type	Municipality
- Mayor	Job Cohen ^[1] (PvdA)
- Aldermen	Lodewijk Asscher Carolien Gehrels Tjeerd Herrema Maarten van Poelgeest Marijke Vos
- Secretary	Erik Gerritsen
Area <small>[2][3]</small>	
- City	219 km ² (84.6 sq mi)
- Land	166 km ² (64.1 sq mi)
- Water	53 km ² (20.5 sq mi)
- Urban	1,003 km ² (387.3 sq mi)
- Metro	1,815 km ² (700.8 sq mi)
Elevation <small>[4]</small>	2 m (7 ft)
Population <small>(1 October 2008)[5][6]</small>	
- City	755,269
- Density	4,459/km ² (11,548.8/sq mi)
- Urban	1,364,422
- Metro	2,158,372
- Demonym	Amsterdamer
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postcodes	1011 – 1109
Area code(s)	020
Website: www.amsterdam.nl	

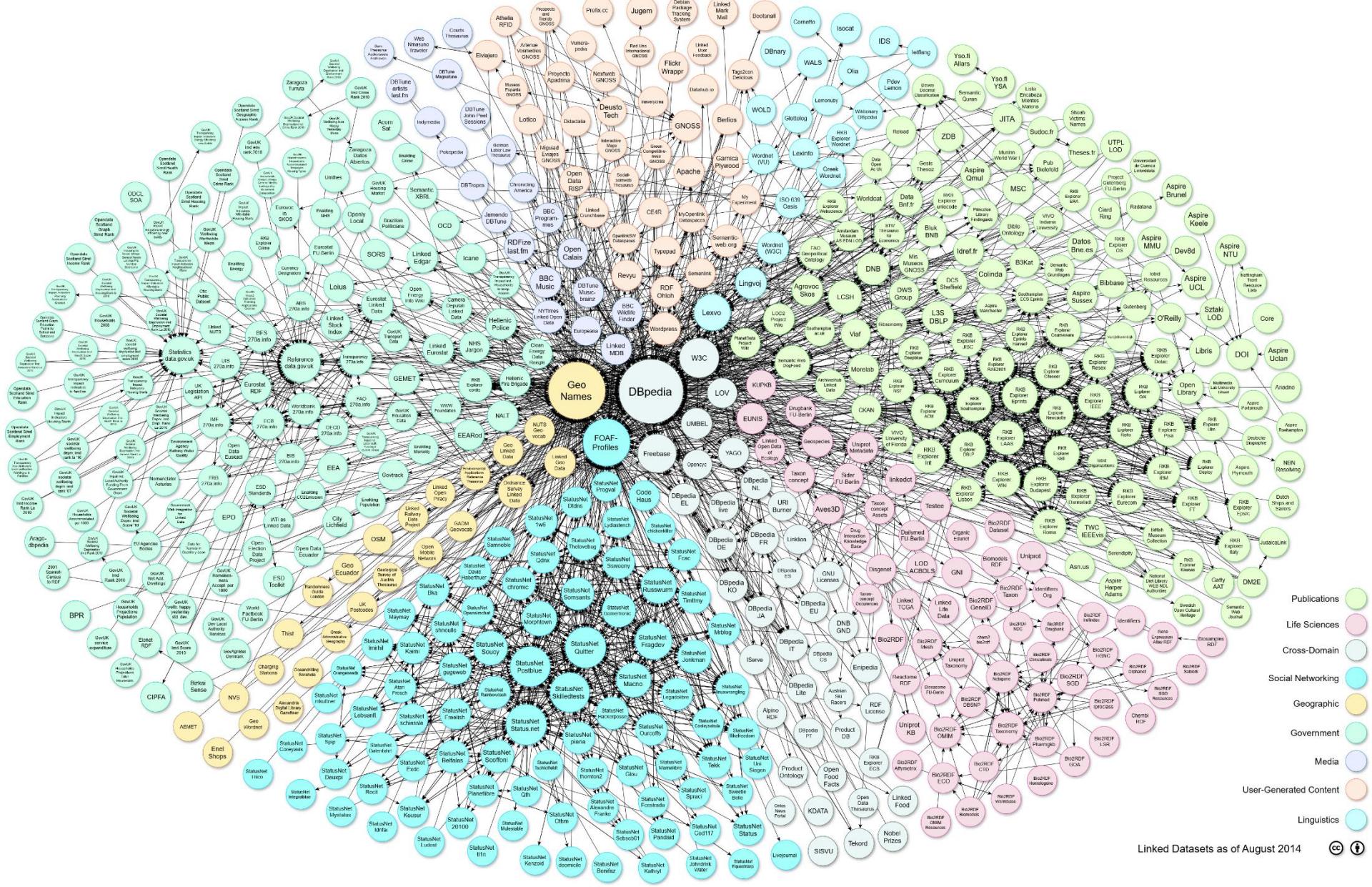
Automatic links among open datasets

```
<http://dbpedia.org/resource/Amsterdam>
owl:sameAs <http://rdf.freebase.com/ns/...> ;
owl:sameAs <http://sws.geonames.org/2759793> ;
...
```

```
<http://sws.geonames.org/2759793>
owl:sameAs <http://dbpedia.org/resource/Amsterdam>
wgs84_pos:lat "52.3666667" ;
wgs84_pos:long "4.8833333";
geo:inCountry <http://www.geonames.org/countries/#NL> ;
...
```

Processors can switch automatically from one to the other...

The LOD “cloud”



View of Knowledge Graphs

Name	Instances	Facts	Types	Relations
DBpedia (English)	4,806,150	176,043,129	735	2,813
YAGO	4,595,906	25,946,870	488,469	77
Freebase	49,947,845	3,041,722,635	26,507	37,781
Wikidata	15,602,060	65,993,797	23,157	1,673
NELL	2,006,896	432,845	285	425
OpenCyc	118,499	2,413,894	45,153	18,526
Google's Knowledge Graph	570,000,000	18,000,000,000	1,500	35,000
Google's Knowledge Vault	45,000,000	271,000,000	1,100	4,469
Yahoo! Knowledge Graph	3,443,743	1,391,054,990	250	800

Table 2.8. Numerical Overview of some Knowledge Graphs, taken from [Paulheim, 2017]¹⁸³.

Case Study: Tirol Knowledge Graph Tourist Domain



Entity	Count
schema: Thing	453,841,147
schema: CreativeWork	175,787,490
schema: MediaObject	175,746,110
http://purl.org/dc/dcmitype/Image	175,735,868
schema: ImageObject	175,735,868
schema: Intangible	172,124,244
schema: StructuredValue	155,482,666
schema: Place	60,996,190
schema: ContactPoint	53,155,166
schema: PostalAddress	51,706,023

Table 2.5. Top 10 entities used in the TKG.

Use of TCG Knowledge Graph

Dialogue-driven room booking:

- Amongst the crawled websites, many are customers of the internet booking engine provider Easybooking
- Developed an Alexa skill that enables voice-driven booking of Easybooking hotels through the TKG.
- If the user asks for a certain hotel, it sends a request to a webhook.
- The result, a list of available hotel offers, is sent back and read to the user.
- The list also contains annotated API descriptions for the booking API.
- If the user decides on an offer, a booking can be executed through a voice command

Use of TCG Knowledge Graph

Showcase dialog system:

- built two dialog systems that fetch their data from the graph.
- One answers generically to touristic topics like hiking or opening hours.
- Other one goes one step further and conducts generic dialogs solely based on data taken from the Knowledge Graph.

Time series analysis of prices in touristic regions:

- prices of offers, if available, are stored permanently, a time series analysis can be conducted.
- price development of two touristic regions can be compared

Time series analysis work perfect with Knowledge Graphs and is a promising application of them in tourism.

Knowledge Graph Key principles

- Findability: The data is identified uniquely and described with metadata following RDF principles.
- Accessibility: the data is accessible over a web UI, through HTTP, or a SPARQL API. Authentication or authorization is not required.
- Interoperability: The language that is used to access the data is SPARQL, which is a W3C recommendation. The vocabulary to describe the data is schema.org, which is a de-facto standard to describe data on the Web.
- Reusable: The data is available openly, the provenance (origin) is stated, and it meets the domain-relevant community standards which were created in cooperation with domain experts.

Bots and Personal Assistants

- Bots and Intelligent Personal Assistants access and aggregate information on behalf of human users.
- First, this requires that these bots understand human language, written or spoken.
- Second, the descriptions of available resources no longer need to be provided to human users directly but must be enriched by semantic annotations to provide machine processability and understandability by bots.
- Both require results from Artificial Intelligence such as Natural Language Processing and Semantic Web Technology facilitating Knowledge Graphs.

Semantic Search and KG

- Google Knowledge Graph, a knowledge base containing already around 100 billion facts about more than 1 billion entities
- Present information from external sources in a structured way become possible
- Only around 25% of the result shows classical web resources
- Google Assistant on mobile devices traditional web results are no longer shown but only content selected and aggregated by Google.
- Alexa, Bixby, Cortana, Facebook messenger, Google Assistant, Siri, and others provide personalized and (spoken) message-based access to information.

Search Results (entity specific)

All Images News Videos Maps More ▾ Tools SafeSearch ▾

About 13,900 results (0.47 seconds)



Chetana Anoop Gavankar :

Overview Videos

 BITS Pilani
<https://www.bits-pilani.ac.in> › pilani › chetana-gavankar

Dr. Chetana Gavankar

Dr. **Chetana** is an Associate Professor in the CSIS department at Work Integrated Learning Programmes Division, BITS Pilani. She has more than 24 years of ...

 LinkedIn · Chetana Gavankar
740+ followers

Chetana Gavankar - Pune, Maharashtra, India

View **Chetana Gavankar's** professional profile on LinkedIn. LinkedIn is the world's largest business network, helping professionals like **Chetana Gavankar** ...

About

Research interests: Natural Language Processing, Machine learning, Deep Learning, Semantic Web

Claim this knowledge panel

Feedback

Profiles

 LinkedIn

 Google Scholar

Query specific UI

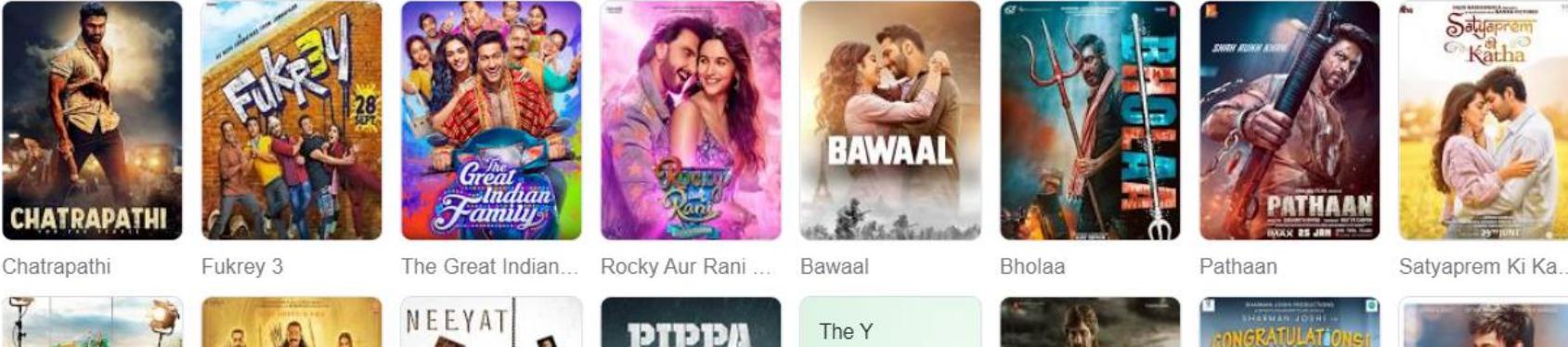
Good movies of 2023

All Images Videos News Shopping More Tools SafeSearch

About 11,45,00,00,000 results (0.62 seconds)

What to watch

Films ▾ 2023 Prime Video Hindi cinema Action Romance Thriller Horror Mystery Prakash Raj Tanishk Bagchi



Chatrapathi Fukrey 3 The Great Indian... Rocky Aur Rani ... Bawaal Bhola Pathaan Satyaprem Ki Ka...

Query specific UI

innovate

achieve

lead

hotels in south g... X Sat, 30 D... < > Sun, 31 D... < > 4

All filters Price Property type Offers Guest rating

South Goa · 2,681 results

The Cape Goa
4.5 ★ (5,294)

5-star hotel Free breakfast Free Wi-Fi
Free parking Hot tub Air conditioning
Beach access Bar Restaurant

View prices

Shree Mangesh temple
Boa Vista Beach ₹5,471
Ponda ₹10,517
Colva Beach ₹6,059
Cavossim ₹9,081
Mobo Beach ₹7,552
Cabo de Rama Fort ₹7,552
Palolem Beach ₹6,359
Netravali Wildlife Sanctuary
Cotigao Wildlife Sanctuary

Update list when map moves

Challenges

- Need to understand user requests precisely
 - Able to organize a goal-oriented dialogue with them.
 - Need to have access to vast sources of smart content, data, and services.
 - Smartness is based on adequately structured and semantically machine processable data.
 - Proper matchmaking of intents identified by Natural Language Processing technologies with semantically annotated resources is needed.
 - Steady improvement and adaptation of achieved solutions through semi-automatic and fully automatic techniques
-

KG Challenges

- Current knowledge graphs fall short on representing time, versioning, probability, fuzziness, context, reification, and handling inconsistency
- Knowledge management technology based on graph-based repositories is responsible for acquiring, storing and managing Knowledge Graphs together with context data on user requests.
- Connection of user request with resources through inference engines based on deductive reasoning

Reasoning

- Reasoning is about discovering new knowledge from existing one.
- It starts with one or more general premises and links them to reach specific conclusions
- If the premises are true, then we conclude that the conclusions are true, too
- While ambiguity can be clarified by using context to infer the intended meaning, become challenging when the knowledge base increases in size.
- Identification and reclassification of misclassified knowledge are also needed, as well as the incorporation of probabilities into reasoning to deal with uncertainty

Reasoning

- Reasoning can be used to find errors, correct errors, and enrich the knowledge in a Knowledge Graph
- Graph database is used to store the Knowledge Graph
- Rules can play different roles in dialog-based systems using Knowledge Graphs.
- Rules allow for expressing complex bearings in an Ontology. For instance, which ski areas are connected via ski slopes, the total length of all ski slopes in an area can be expressed by rules.
- Reasoning dynamically deduces the answers using those rules

Reasoning

- Reasoning uses the context, the previous answers, and the Knowledge Graph to create flexible and intelligent dialogs.
- Reasoning allows integrating external knowledge sources and services context into the answers
- Requirements are high performance and scalable real-time reasoning

Knowledge access layer advantages

- Access layer can provide a reduced amount of data
- Not necessary to curate the entire Knowledge Graphs in terms of aspects like inconsistencies.
- Semantic data lake allow different views on data.
- Each reasoning agent can restrict itself to a well-curated island.
- Access rights and privacy issues may prevent specific applications to access the entire Knowledge Graph but only subsets of it. This can be implemented by these views.
- It integrates additional dynamic knowledge sources and user given input

Knowledge access layer

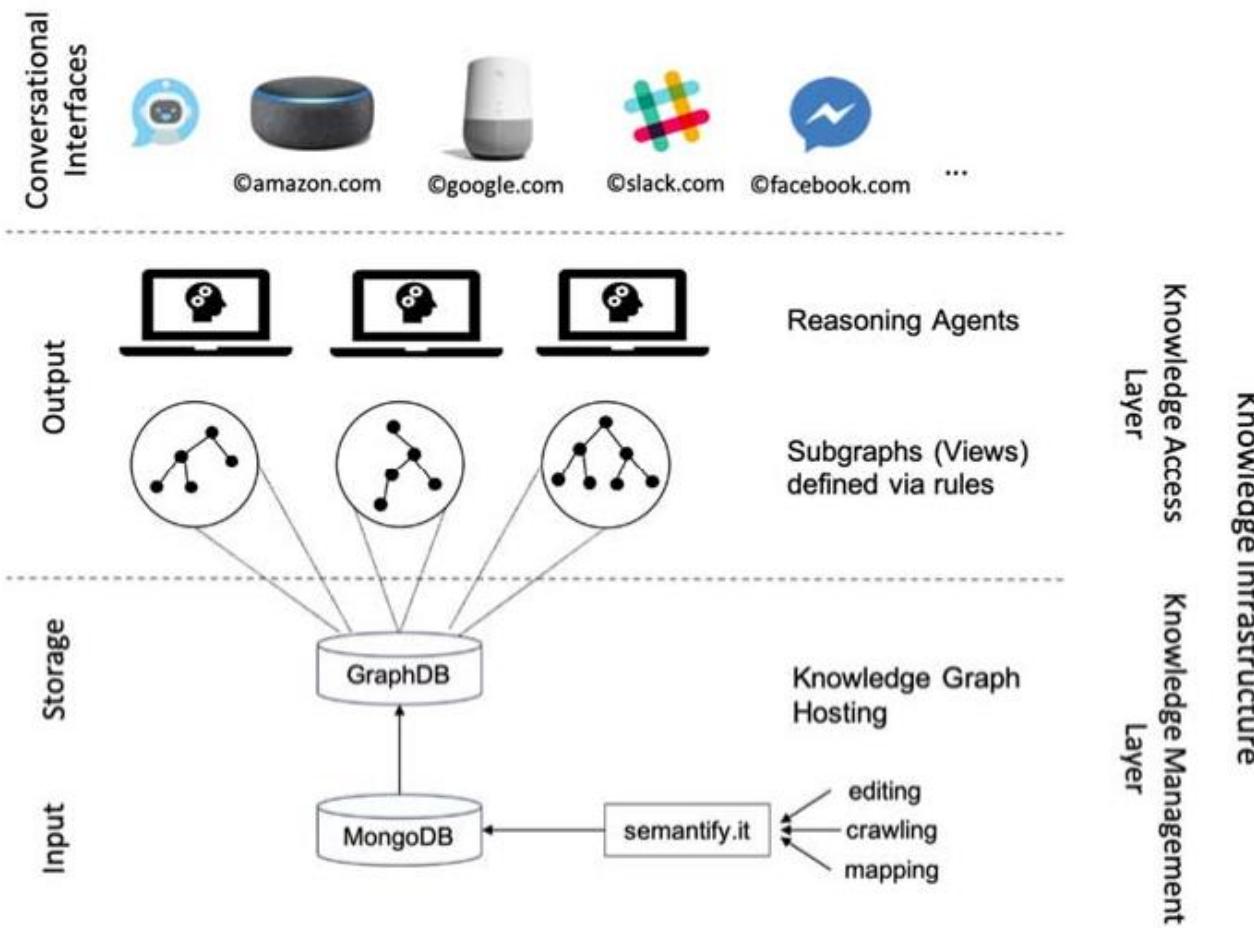


Fig. 3.8. Reasoning agents accessing a Knowledge Graph.

Knowledge access layer

Applications

- Semantic Search
- Question Answering
- Analytics
- Dashboards
- Knowledge Sharing
- Knowledge Management

Algorithms

- Inferencing
- Machine Learning
- Entity Recognition
- Disambiguation
- Text Understanding
- Recommendations

Knowledge Graph



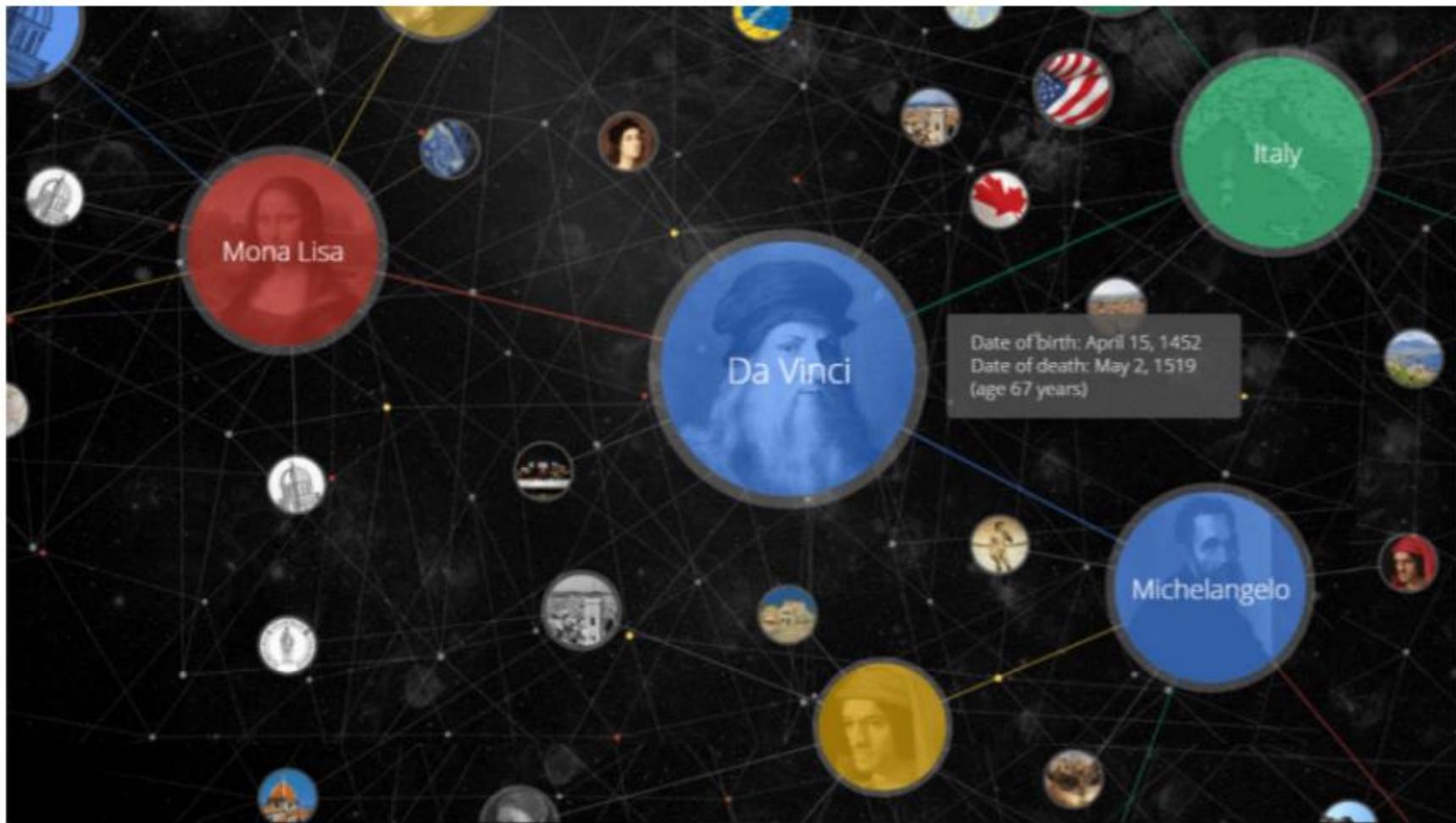
- Entities
- Relationships
- Semantic Descriptions

Data Sources

Data Transformation, Integration
Natural Language Processing



Google Knowledge Graph



Semantic Search

Renewable Energy Glossary

- energy (12)
- wind (7)
- windpower (6)
- wind farms (5)
- wind turbines (5)
- generators (4)
- photovoltaic power (4)
- ocean energy (2)

[More](#)

Climate Compatible Development Glossary

- renewable energies (6)
- wind power (6)
- projects (4)
- economic cost (3)
- fossil energy (3)
- greenhouse gas emissions (3)

wind farms

 Show tagged content only

Wind farms



A wind farm is a group of wind turbines in the same location used for production of ... For example, Gansu Wind Farm, the largest wind farm in the world, has several thousand turbines ...

[WIND FARMS](#)

[SHOW PARENT](#)

Basics of Wind Energy



What is a wind farm ... The turbines in a wind farm are connected so the electricity they generate can travel from the wind farm to the power grid ...

[ELECTRICITY GENERATION](#) [ENERGY](#) [GRIDS](#) [WIND](#) [WIND POWER](#) [WIND TURBINES](#) [WINDPOWER](#)

[SHOW PARENT](#)

Wind_test



Wind farms consist of many individual wind turbines which are connected to the electric ... Small onshore wind farms can feed some energy into the grid or provide electricity to ...

[ECONOMIC COST](#) [ELECTRICITY GENERATION](#) [ENERGY](#) [WIND](#) [WIND FARMS](#) [WIND POWER](#) [WIND TURBINES](#) [WINDPOWER](#)

[SHOW PARENT](#)

France Gears Up for Floating Wind



This project intends to quickly put floating wind farms on the map as a competitive energy ... Engie has stated: "Floating wind turbines are an up-and-coming technology that can be ...

[SHOW PARENT](#)

[SHOW PARENT](#)

wind farms



Description

A group of wind turbines interconnected to a common power provider system through a system of transformers, distribution lines, and (usually) one substation. Operation, control, and maintenance functions are often centralized through a network of computerized monitoring systems, supplemented by visual inspection.

Also known as

wind parks, wind power plants, wind power stations, windfarm

More general

windpower concepts

More specific

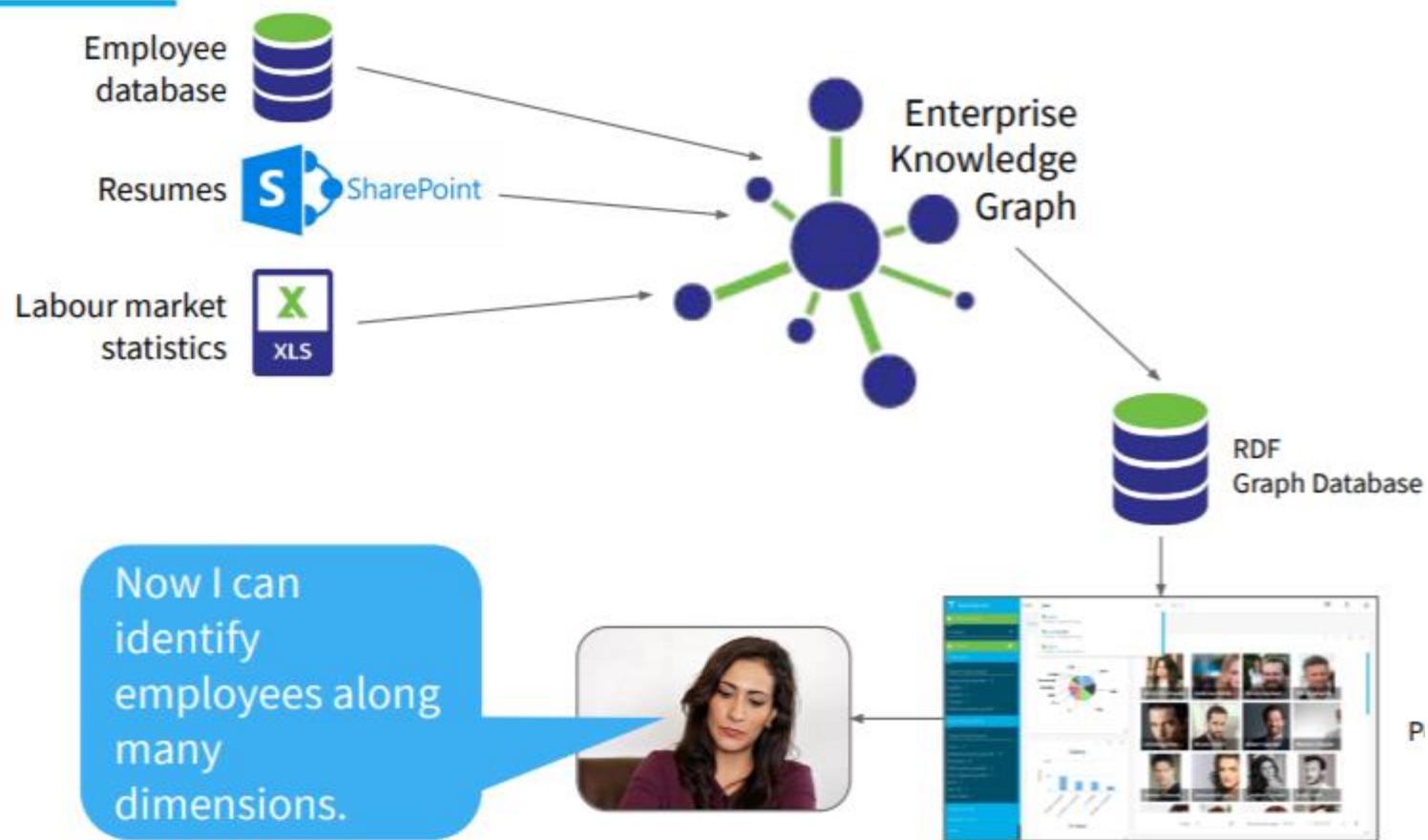
grid-connected wind power systems, on-shore windparks, off-shore windparks

Related

wind turbines, wind power capacity installed, wind feed-in tariffs, stand-alone wind power turbines

Knowledge Graphs for **Data Integration & Analytics**

Metadata enrichment, linked data, text mining, entity-centric search, agile reporting



Research in Life Sciences

As a researcher in pharmaceutical industry, I want to plan new experiments more efficiently. I want to know what's already available. I'm interested in former experiments where

- certain genes were tested
- under specific treatment conditions
- in a target therapeutic area
- with help from categorisation systems like 'disease hierarchies'

→ Linking Structured to Unstructured Data and to Industry Knowledge Graphs

UniProt, ChEMBL



Experiments Documentation



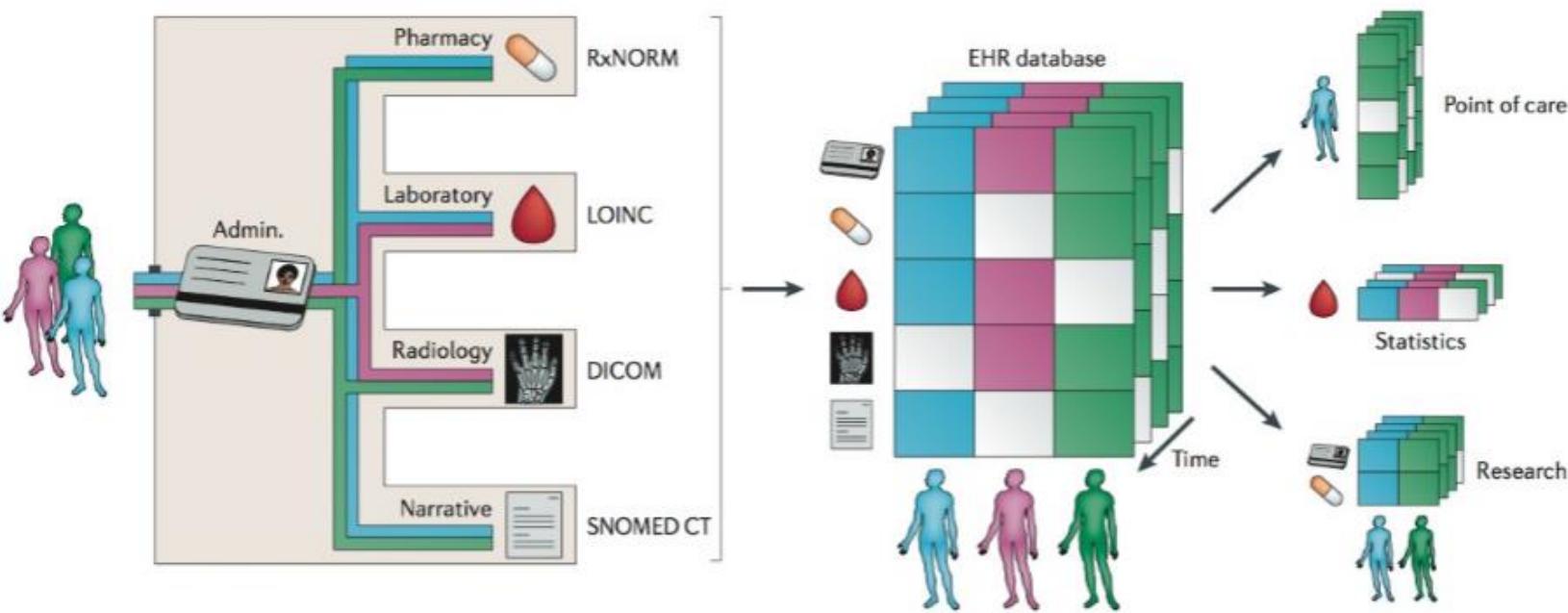
MeSH



DrugBank

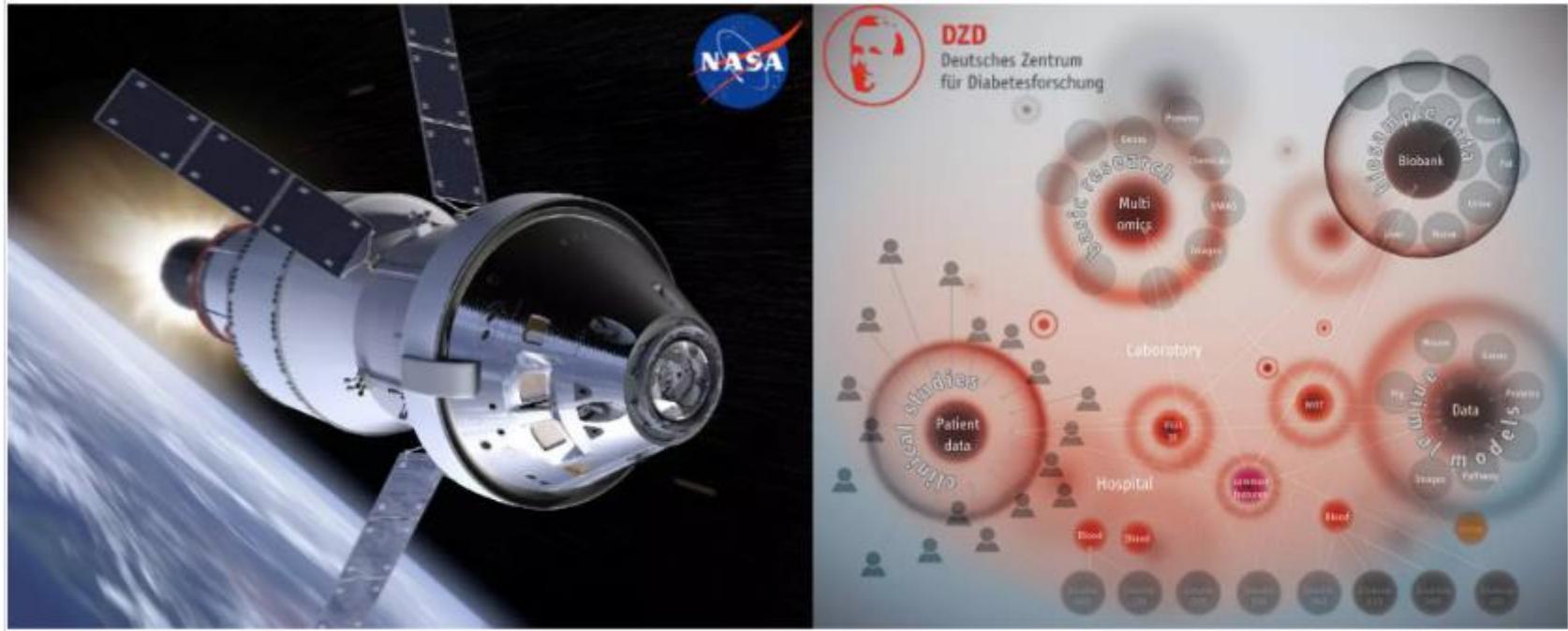


Data and text mining of electronic health records



<https://www.slideshare.net/larsjuhljensen/data-and-text-mining-of-electronic-health-records-55027555>

Real world use of KG



Real world use of KG



Types of Data

- Knowledge Graphs and applications working with them deal with three different types of data:
 - **Static data** are data that do not likely change frequently. For instance, the address of a hotel or the birthdate of a person do not change very often. This low velocity allows to directly store these data in a Knowledge Graph repository.
 - **Dynamic data** change with a high frequency, for instance, age, weather forecast, traffic information or share prices. Here the actual data must be accessed periodically or on demand. It is possible to store these data for future data mining, but the current data must be accessed dynamically.
 - **Active data**, which changing the state of an external resource. For instance, a booking action changes the state of an external resource (e.g. creation of new room reservations at a hotel)

Semantic Web Services

- For dynamic data like temperature at a given location and time, a request (e.g. HTTP GET) to the weather web service must be made to retrieve the necessary information at the query-time.
- For active data, a request (e.g. HTTP PUT) to a hotel's booking web service must be sent to make a room reservation.
- Intelligent applications like conversational interfaces then access the web services on the fly

Challenges in QA

Lexical Gap: Vocabulary used in the question does not directly match the vocabulary used in the labels of the resources in a Knowledge Graph

Ambiguity: notion of the same phrase having different meanings. benefit from the statistical distribution of phrases in several text corpora to identify the correct context of a question, may benefit from the graph structure of Knowledge Graphs in order to rank the relatedness of entities. Some moved the effort mostly to the user and let them do the disambiguation of candidate answers

Challenges in QA

Multilingualism

- refers to the notion of being able to handle questions and answers in multiple languages.
- Open Knowledge Graphs such as DBpedia and Wikidata have mappings of the data in different languages.
- Such mappings make a precious resource for developing multilingual QAS.
- use multilingual lexicon Ontologies for answering multilingual questions

Complex Questions and Queries:

- Can benefit from pre-built linguistic representations of concepts and relations in Ontologies, determining the core elements and building queries around it, even from multiple sentences
- handle such questions by inferring such temporal and spatial relationships based on the existing knowledge

Knowledge Graphs and Dialog systems



- Two main directions in which one can use a Knowledge Graphs for this purpose:
 - to power the language understanding part of the dialogue system and
 - to react on the conversations and provide additional interactions, information, and recommendations to the user engaged in conversations with the dialog system

Language Understanding

- Use the Knowledge Graphs to provide training data for the Natural Language Understanding service (e.g., DialogFlow).
- Ingest from the Knowledge Graph as training data for the entity recognition task (e.g., Vienna is a City) and provide (semi-) automatically generate intents and example questions.
- Based on the Knowledge Graph structure
 - we can generate on the one hand entities and synonyms
 - intents needed in the Natural Language Understanding service based on the entities,
 - the relations between these entities in the Knowledge Graph
- ontology-to-text approaches to generate example questions that can be used to train the Natural Language Understanding service

React on the conversations

- Can elaborate on the topic of discussion and provide additional interesting facts.
 - Let us consider a dialogue system providing information about events.
 - Based on the knowledge from a touristic Knowledge Graph, system can provide the user with detailed information about an event, but also additional information about the connected entities, e.g., artist, location, and means of transportation related to that event
- Improve the handling of the conversation context
 - Using a template or rule-based approach one can drive interaction with the user to provide context-dependent answers (e.g., prices based on the role of the user), the invocation of actions (booking of a hotel room with follow up questions), or a combination of both (e.g. which outdoor events happen on the weekend if the weather permits).
- Refine the search for products or services in a dialogue system
 - actions and APIs that are modelled as part of the Knowledge Graph might require input data of a specific type, and can be used to steer the dialogue

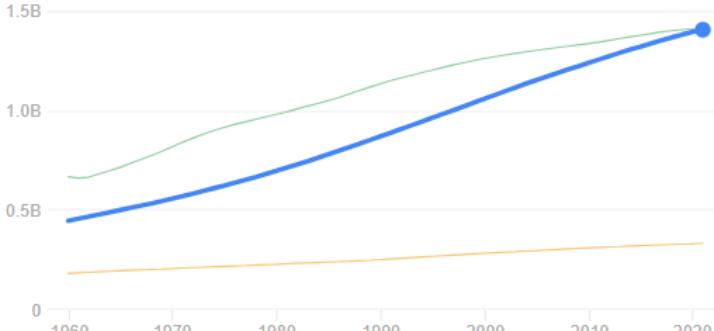
QA Results using Knowledge Graph

Google search results for "what is population of india":

About 1,15,00,00,000 results (0.96 seconds)

India / Population

140.76 crores (2021)



Country	Population (2021)
China	141.24 crores
India	140.76 crores
United States	33.19 crores

India
Country in South Asia

Related statistics

Gross domestic product	3.18 lakh crores USD (2021)
Population growth	0.8% annual change

Semantic Annotation and Dialog Systems

- Dialogue system can generate a backend driven intent and training sentences semi-automatically.
- It can benefit from the hypermedia links of a Web API to extract the invocation flow that guides the dialogue

User: *I want a hotel room in Seefeld from 12.09.2019 to 14.09.2019*

Bot: *How many guests?*

User: 2

Bot: *Ok I have found following hotel rooms: 1. Double room, Shower or Bathtub, WC offered by Alpenpark Resort for 200 EUR per night. 2. Double room, shower or bath, toilet offered by Central Seefeld for 119 EUR per night. You can purchase an offer.*

User: *Ok I want number 2. . . .*

.....

Bot: *Done. Here is the reservation:*

Semantic Annotation and Dialog Systems

- First the action is processed and a search intent for lodging reservation including its required parameters is created.
- Then user's request that already contains some of the required information for booking a room is matched with this intent.
- After the missing input is elicited (i.e. occupancy), the search is done via the endpoint defined in the action.
- The action returns hotel rooms with offers and potential buy actions attached to them.
- After user selects an offer, the buy action on that offer is completed and reservation is shown to the user.

LLM

LLMs Give Us an Amazing Opportunity to:

1 Automate data retrieval tasks

2 Improve customer service experiences

3 Expedite reading, understanding, & summarizing

4 Generate content & code

LLM Challenges

But There Are Challenges...

1 Knowledge cut-off

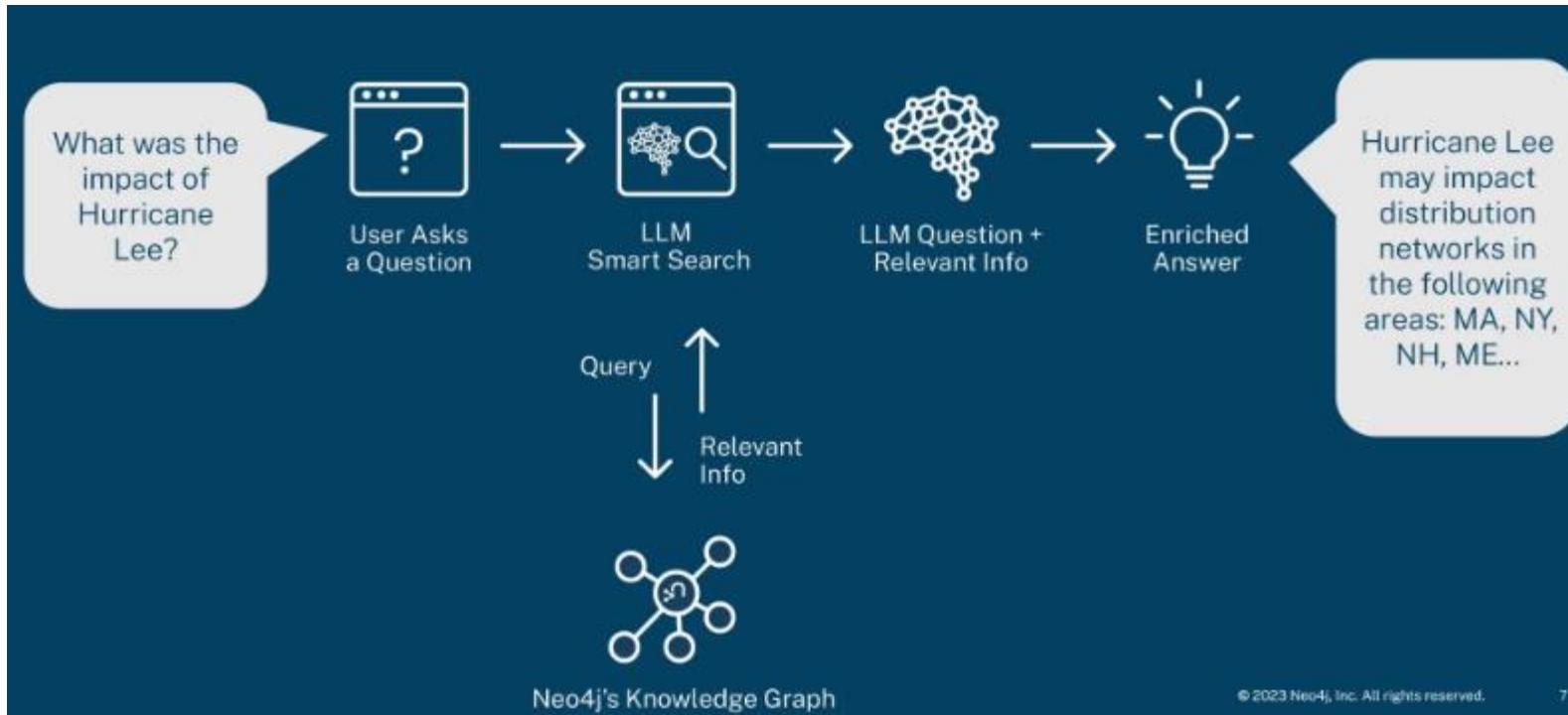
2 Reasonable answers, not always accurate

3 Can inherit bias through training data

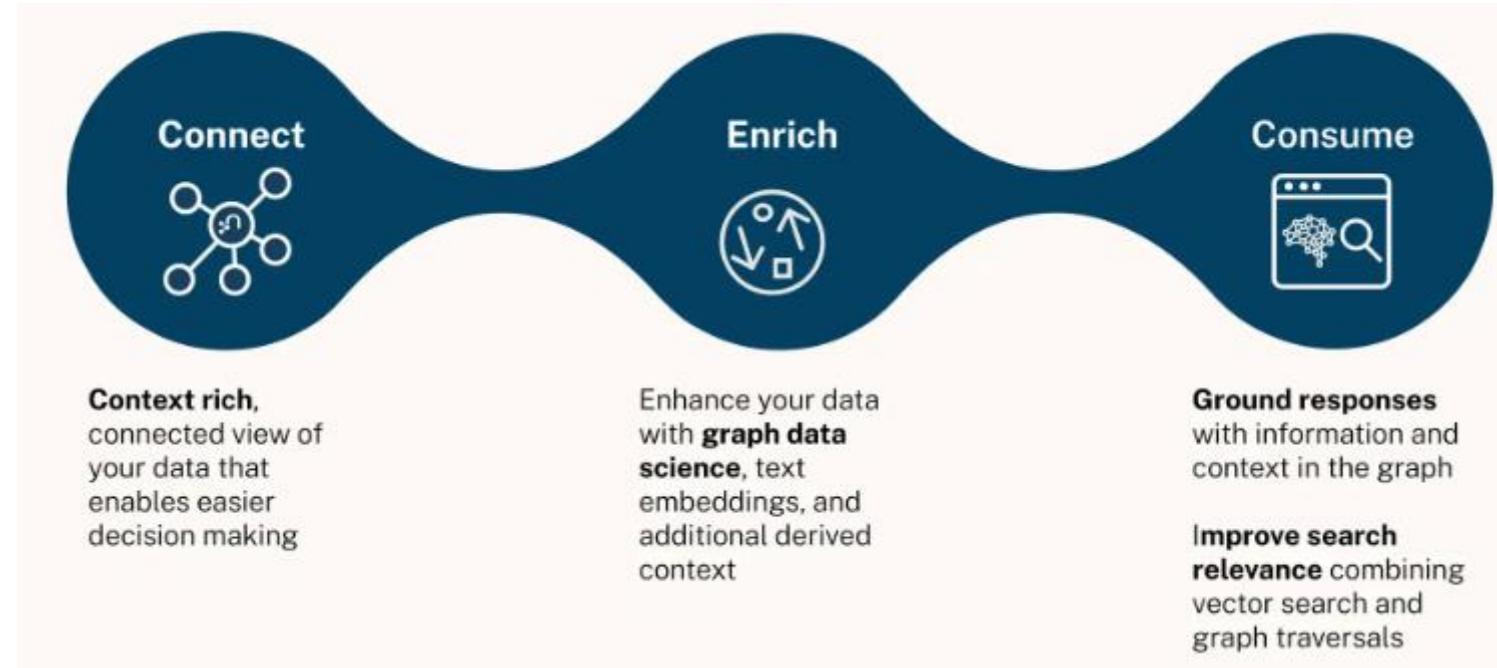
4 Lack of enterprise domain knowledge

5 Inability to verify or attribute sources

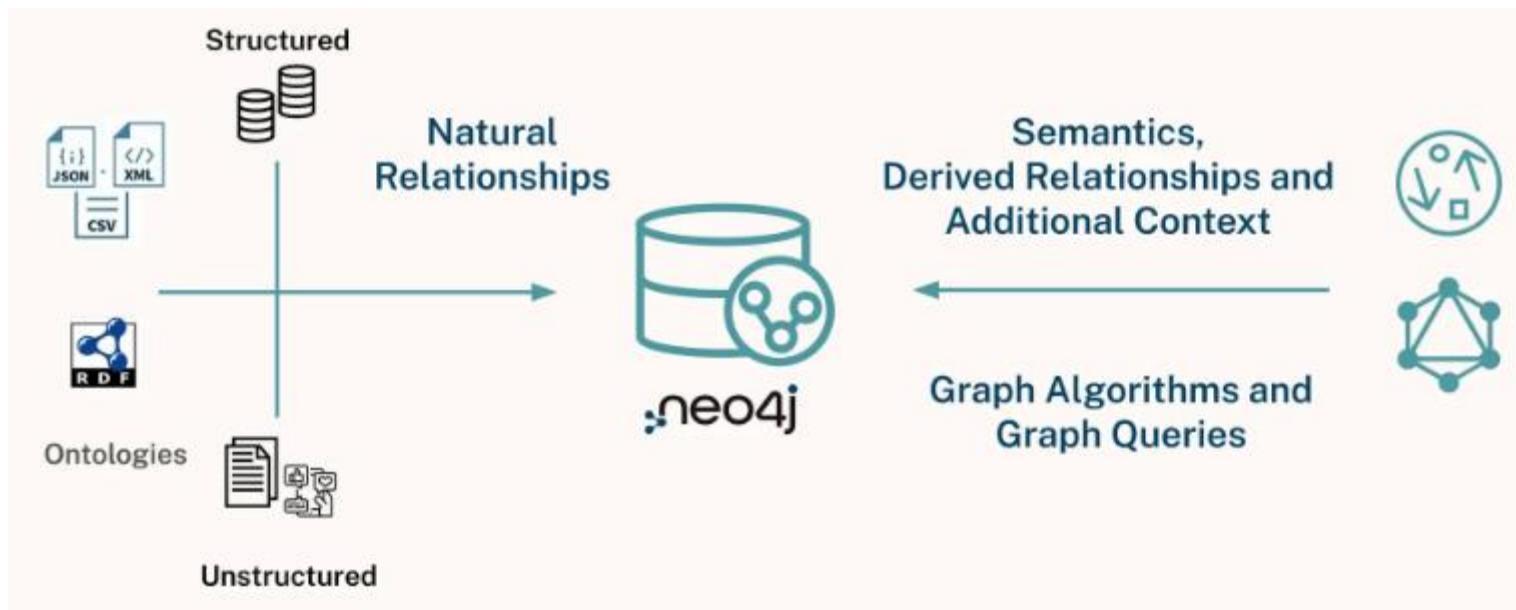
Knowledge Graph use to ground LLM



Knowledge Graph use to ground LLM



Knowledge Graph



Knowledge Graph Use

Finance

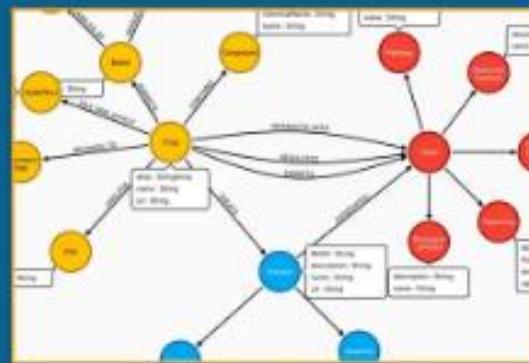


How many flagged accounts are in the applicant's network **4+ hops out?**

How many **login / account variables in common?**

Add these metrics to your approval process

Life Sciences



What **completes the connections** from genes to diseases to targets?

What genes can be reached **4+ hops out** from a known drug target?

What **mechanisms in common** are there between two drugs?

Marketing & Recommendations



Collaborative filtering: users who bought X, also bought Y

What items make you more likely to buy additional items **in subsequent transactions?**

Traverse hierarchies - what items are similar **4+ hops out?**

Semantic Search

Text Embedding Vectors for Semantic Search

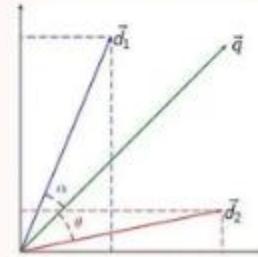
graphsun [

Given a question, find the most relevant documents based on a similarity metric (such as Cosine Similarity) between vector of the question and vectors of contents.

Moving from keyword search to similarity (semantic) search.

Q: what is a text embedding?

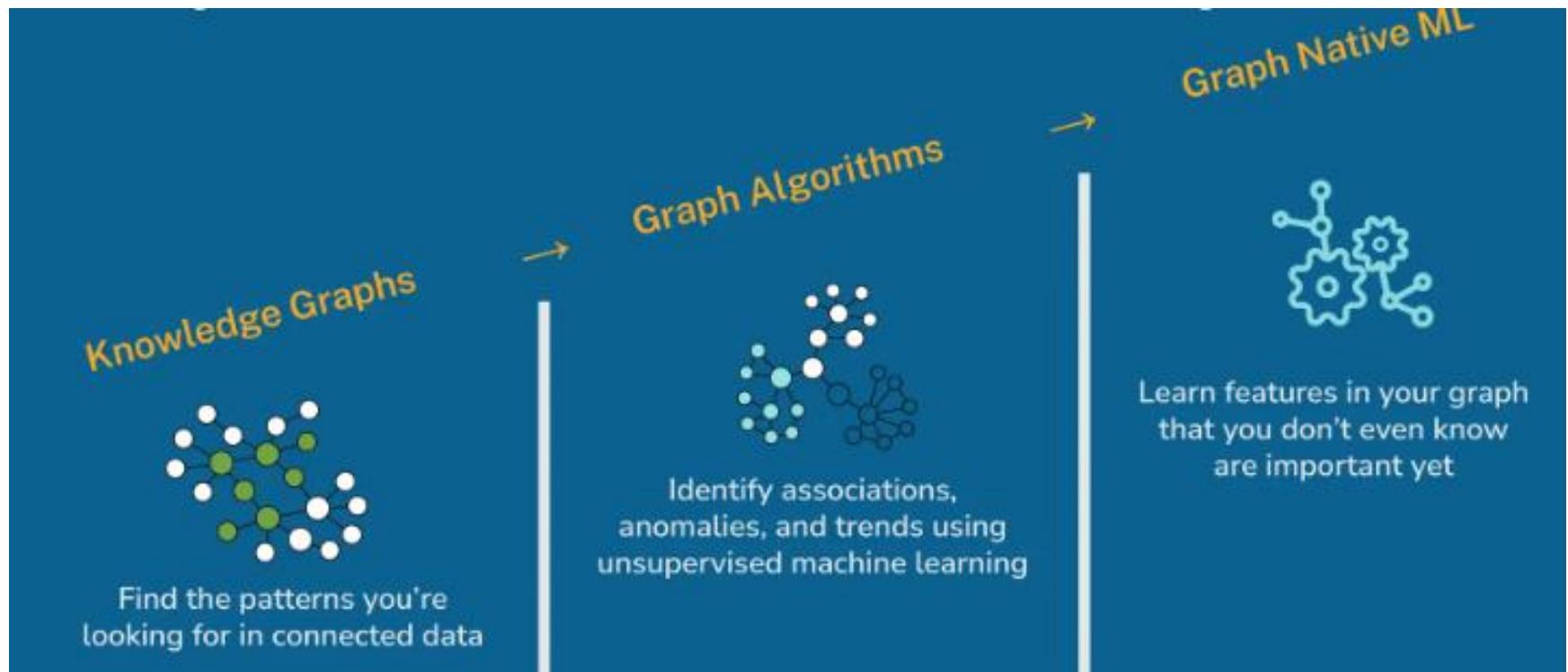
embedding [-0.019200351,0.0035306285,0.0
14364655,-0.009949144,0.00115
81815,0.017021084,-0.01998375,
-0.018986698,-0.033643346,-0.0
2841595,0.014685135,0.0141937
32,... Show all]



Top K by similarity

abstractId	similarity
456	0.923445
22	0.892114
...	...

KG and ML/DL

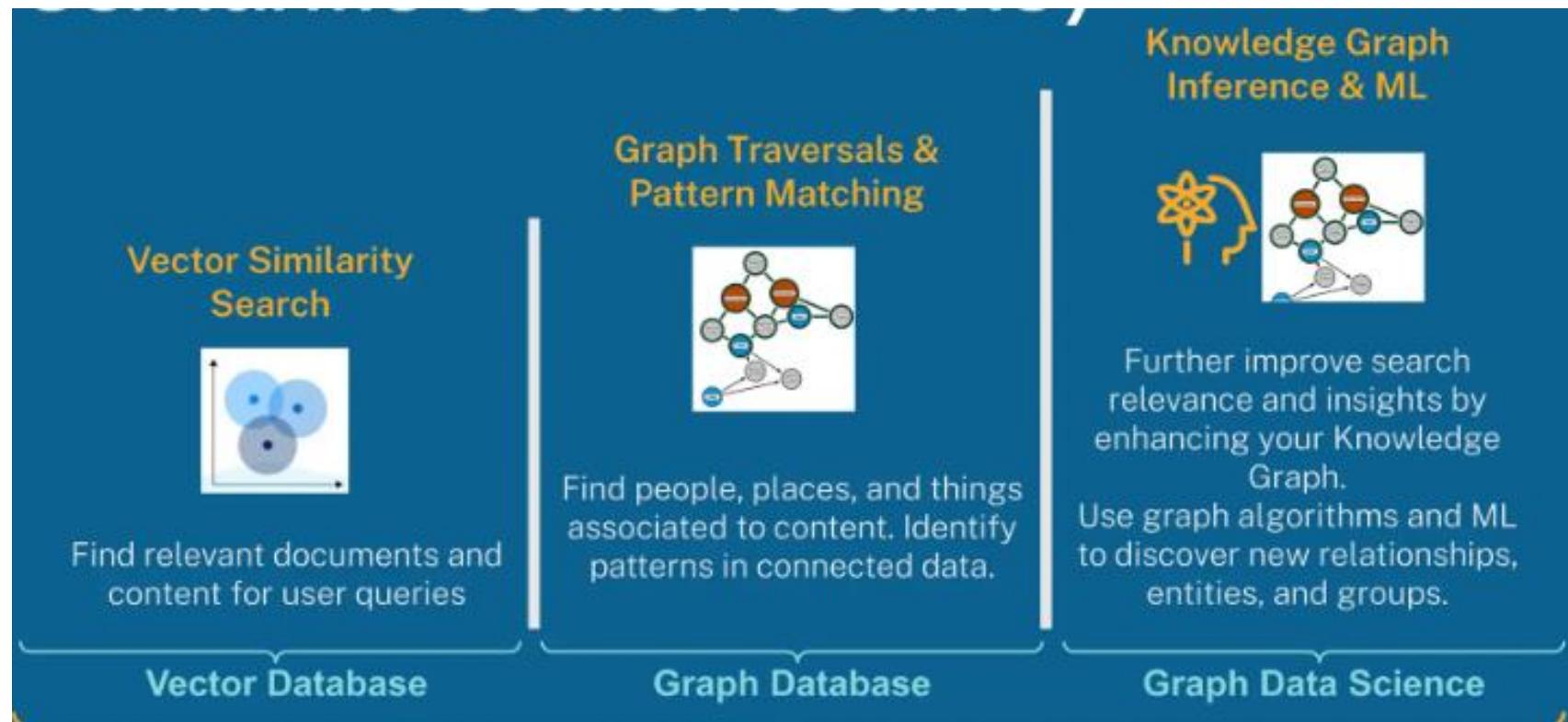


Semantic Search

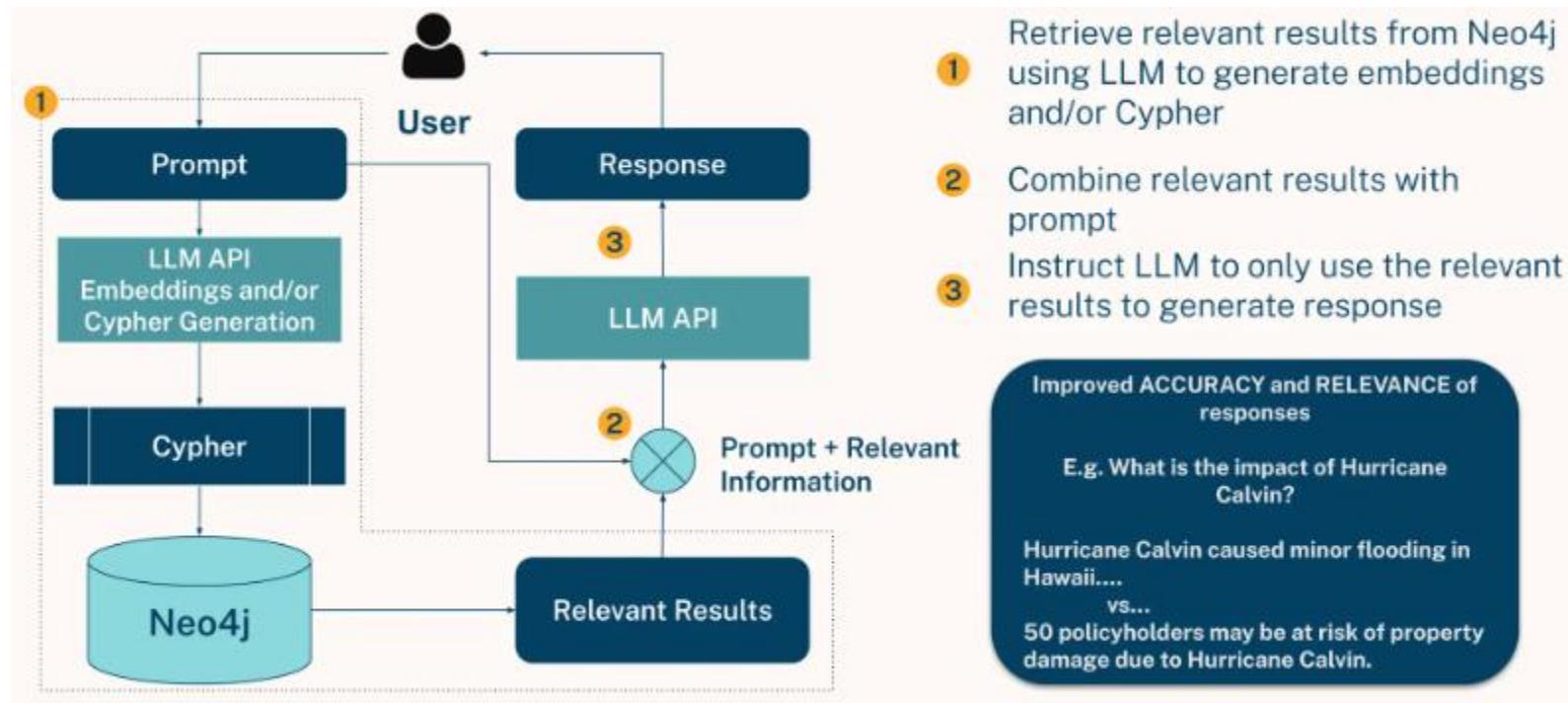
innovate

achieve

lead



Improve results with knowledge graph using Neo4j

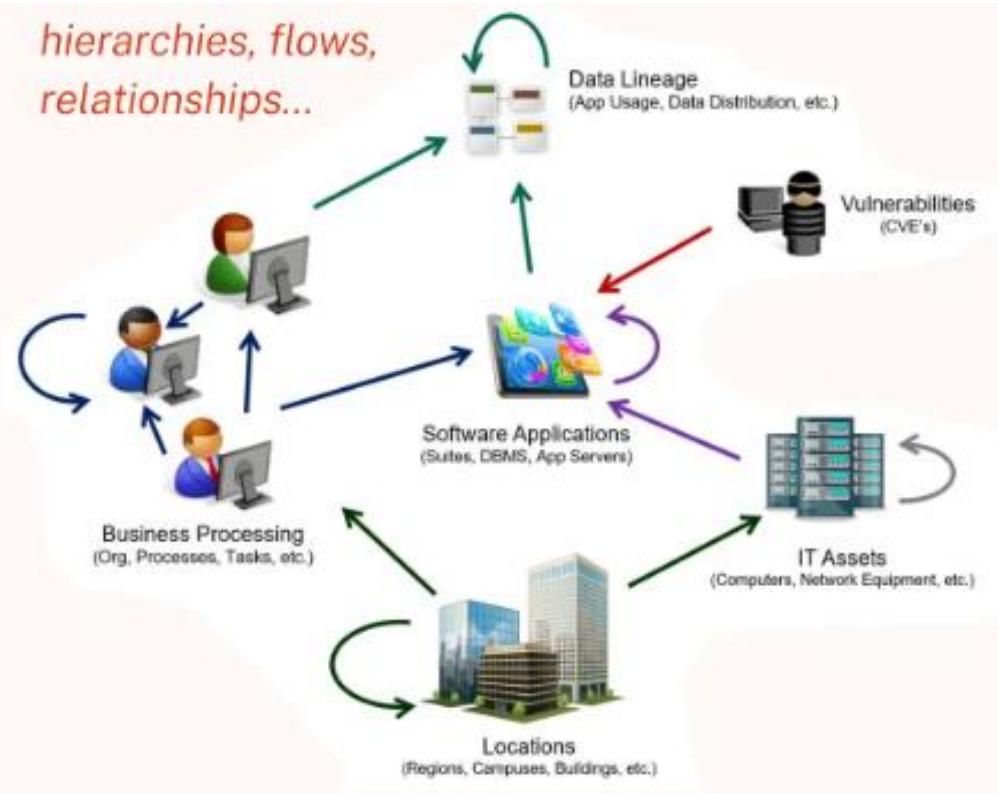


Case Study

Analyze business impact of

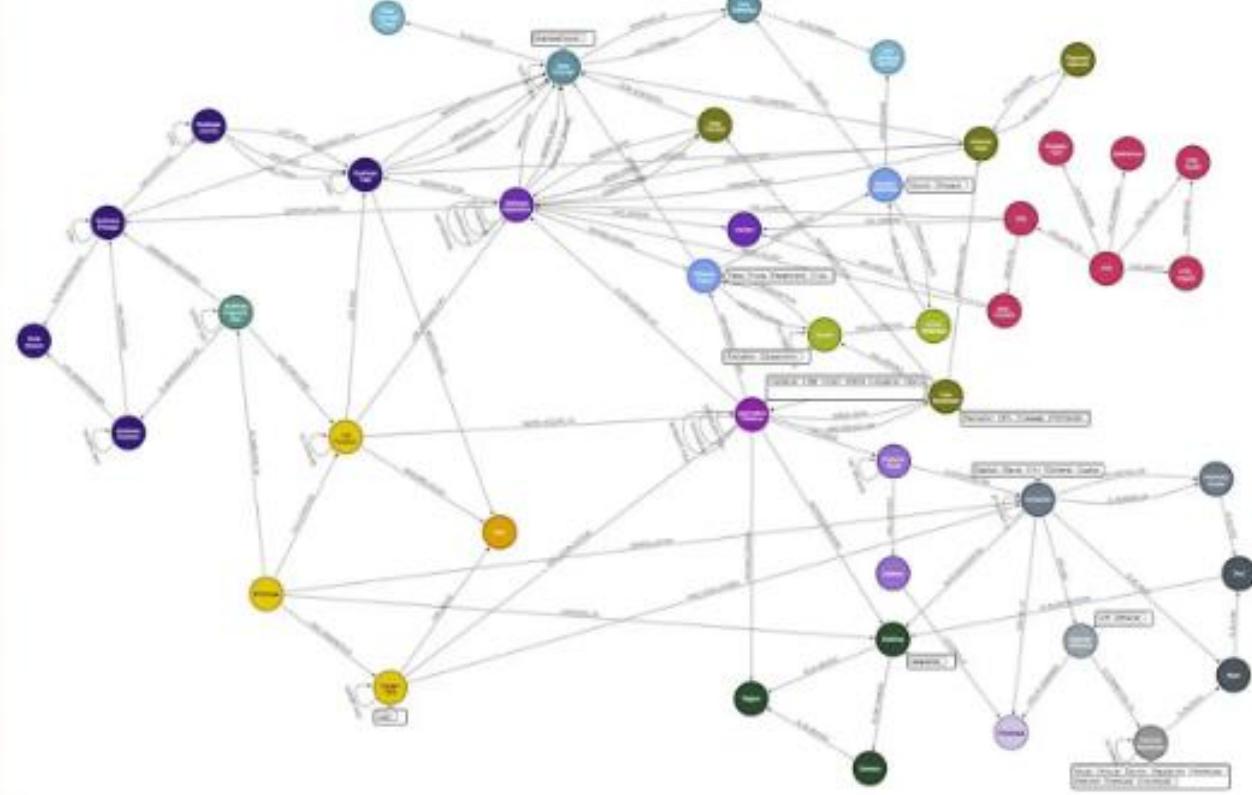
- software & OS vulnerabilities,
- hardware & software upgrades,
- building/geographic disasters
- changes to business data formats

...across mission critical applications and business locations

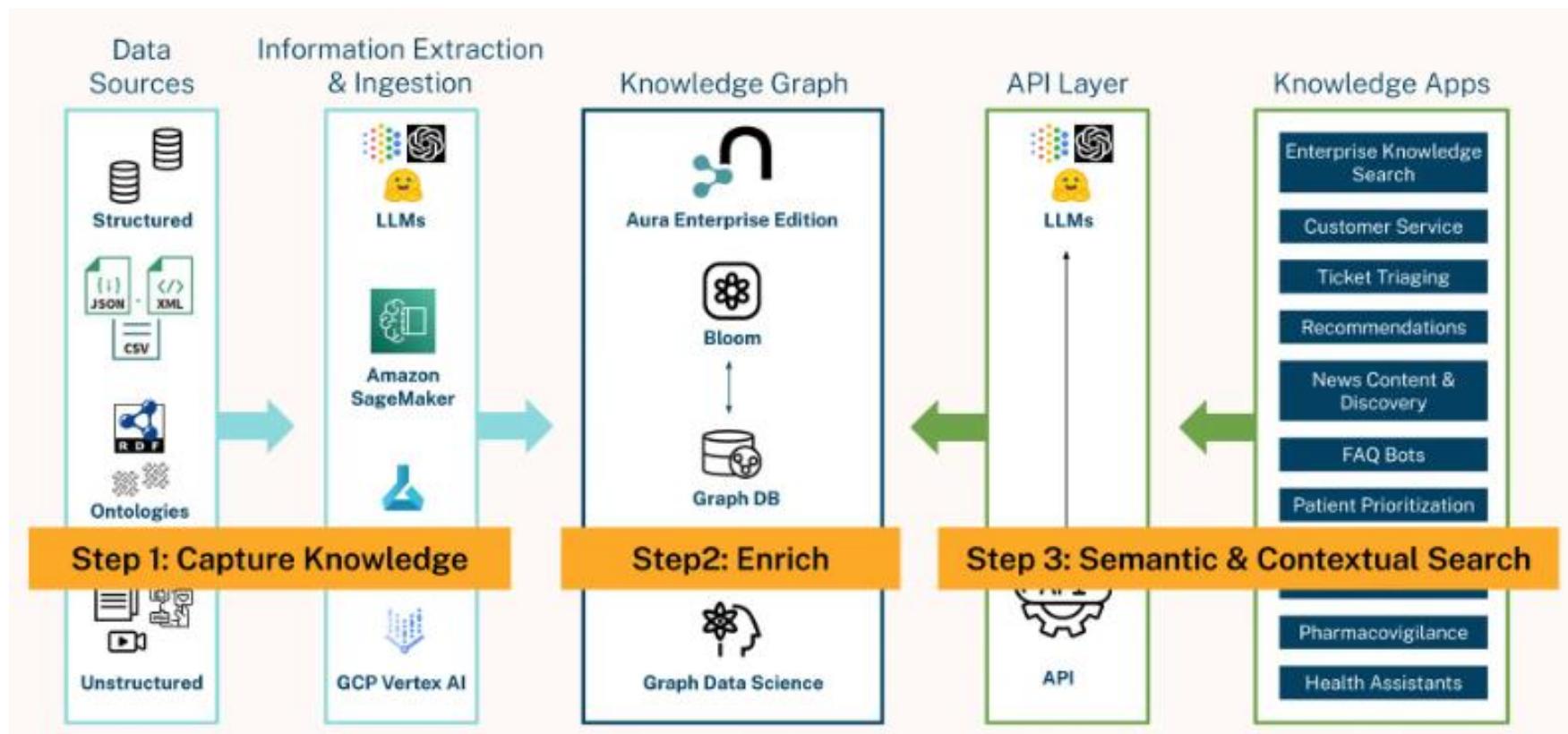


Data Model

CVE Data
Business Data Elements
Vendors, Software
Business Tasks
Application Instances
Data Transfers
People/Roles
Locations
IT Assets



Use of KG



LLM and knowledge graph

- Text embedding generated by a LLM can encode any text (unstructured) into a fixed size list of floats (vector), which can then be matched by using one of the Similarity algorithms.
 - integrated approach that combines LLM with KG can understand natural language queries and fetch information regardless of its underlying structure, leading to a smoother and more intuitive search experience
 - *generate a Cypher query from a question in natural language*, using GPT-3 and other alike LLMs.
 - Query is generated by LLM based on examples provided, which is then executed by the underlying database to return results.
-

Challenges using LLM

- LLM can't always generate correct label, relationship or property names, even if the schema is given as part of the *prompt*
- LLM can't generate complex query if there is no sample given
- Due to prompt size limit, it's not possible to provide enough samples so LLM may not be able to learn all available syntax of the query language
- LLM can mess up with relationship directions

LLM and Knowledge Graphs

- LLMs challenges
 - Coherence
 - Provenance
 - Semantics
 - Explainability
- Knowledge Graph challenges
 - Time to build
 - Expertise requirements
 - Scalability
 - Integration

LLM and Knowledge Graph collaboration



- Transparency / Provenance
- Accuracy
- Privacy
- Fairness
- Accountability

LLM and Knowledge Graph

1. LLMs accomplishments should be acknowledged & leveraged
2. KGs may help communicate & interoperate
3. KGs may help large language models move up the Data-Information-Knowledge-Wisdom pyramid
4. KGs may help address policy and accountability needs/requirements
5. KGs and LLMs *could* be a complementary “marriage”
6. COLLABORATE!
7. Explore opportunities, create exploration “sandboxes”
8. Assessment tools are potentially a BIG opportunity
9. Harmonized data portals are still needed
10. Hybrid AI solutions and pipelines are an opportunity
and...

Few domains KG is used

Bioportal – medical
Government
Cyber security
Smart City – Urban Data Analytics
Navigation
Amazon-Usage, movies, series, trends etc
Weather
Financial Services
Manufacturing- supply chain

References

- Berners-Lee, Hendler, Lassila (May 17, 2001). The Semantic Web. Scientific American Magazine. @ <http://bit.ly/2z2LZIE>
- Berners-Lee. A roadmap to the Semantic Web. Tim Berners-Lee's Design Issues @ <http://bit.ly/2z29fGP>
- Berners-Lee. Linked Data. Tim Berners-Lee's Design Issues @ <http://bit.ly/21MR3Zt>
- Heath, Bizer. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology @ <http://bit.ly/2xIWKou>
- <https://wiki.dbpedia.org/services-resources/ontology>
- <https://www.emse.fr/~zimmermann/Teaching/SemWeb/semwebintro.pdf>
- <https://www.youtube.com/watch?v=C9M7N979pIU>
- <https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/>
- <https://www.w3.org/Consortium/facts>
- <https://pages.semanticscholar.org/coronavirus-research>

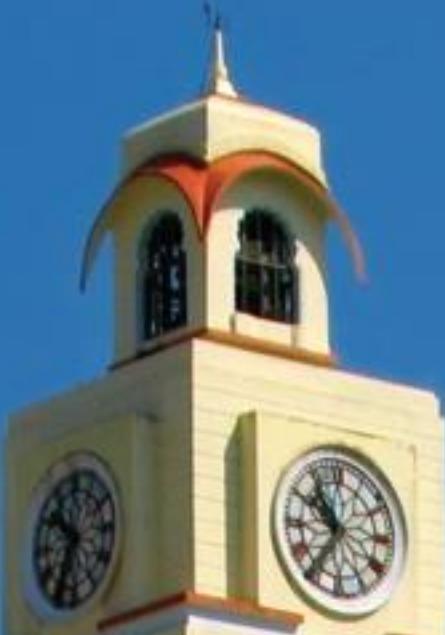
References

- <https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph/>
- [Google Knowledge Graph](#)
- [Building Knowledge graph](#)
- <https://www.youtube.com/watch?v=lUc0woFX16M>
- https://www2.slideshare.net/jeffpan_sw/linked-data-and-knowledge-graphs-constructing-and-understanding-knowledge-graphs?qid=ab5124c7-a15a-4f77-9fc9-914feac00a6a&v=&b=&from_search=1
- https://www2.slideshare.net/phaase/getting-started-with-knowledge-graphs?from_action=save
- <https://lod-cloud.net/>
- <https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/>
- https://www.slideshare.net/slideshow/embed_code/key/dJaP9LmSd1uPXq
- <https://www.youtube.com/watch?v=hILYM9oBI0M>
- <https://www.youtube.com/watch?v=sYejXeuZaeY>

References

- <https://www.slideshare.net/STI-Innsbruck/knowledge-graphs-236706310>
- <https://www.slideshare.net/TakanoriUgai/practical-use-of-knowledge-graph-with-case-studies-using-semantic-web-publishing-tools>
- <https://www.slideshare.net/elenasimperl/knowledge-graph-use-cases-in-natural-language-generation>
- <https://redirect.cs.umbc.edu/courses/graduate/691/fall18/07/resources.html>
- <https://towardsdatascience.com/analyze-your-website-with-nlp-and-knowledge-graphs-88e291f6cbf4>
- https://ckg.readthedocs.io/en/latest/advanced_features/c_kg-notebooks.html

- <https://towardsdatascience.com/analyze-your-website-with-nlp-and-knowledge-graphs-88e291f6cbf4>
- https://colab.research.google.com/drive/1QUpHvcFvNsWyRZM_J5ahsLhEHJY4SjyS#scrollTo=P5rpiKd9Wbjc
- <https://neo4j.com/developer-blog/knowledge-graph-based-chatbot-with-gpt-3-and-neo4j/>
- https://www.youtube.com/watch?v=vA2vd_eUglo
- [https://medium.com/@yu-joshua/future-of-knowledge-graph-will-structured-and-semantic-search-come-into-one-952d33951df3#:~:text=Knowledge%20graphs%20\(KGs\)%20have%20roots,or%20so%20years\(already!\).](https://medium.com/@yu-joshua/future-of-knowledge-graph-will-structured-and-semantic-search-come-into-one-952d33951df3#:~:text=Knowledge%20graphs%20(KGs)%20have%20roots,or%20so%20years(already!).)



Natural Language Processing Applications



BITS Pilani
Pilani Campus

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 7: Knowledge Graph Applications

Date – 7th January 2024

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Philip Cimiano , Johanna Völker , Paul Buitelaar and many others who made their course materials freely available online.

Session Content

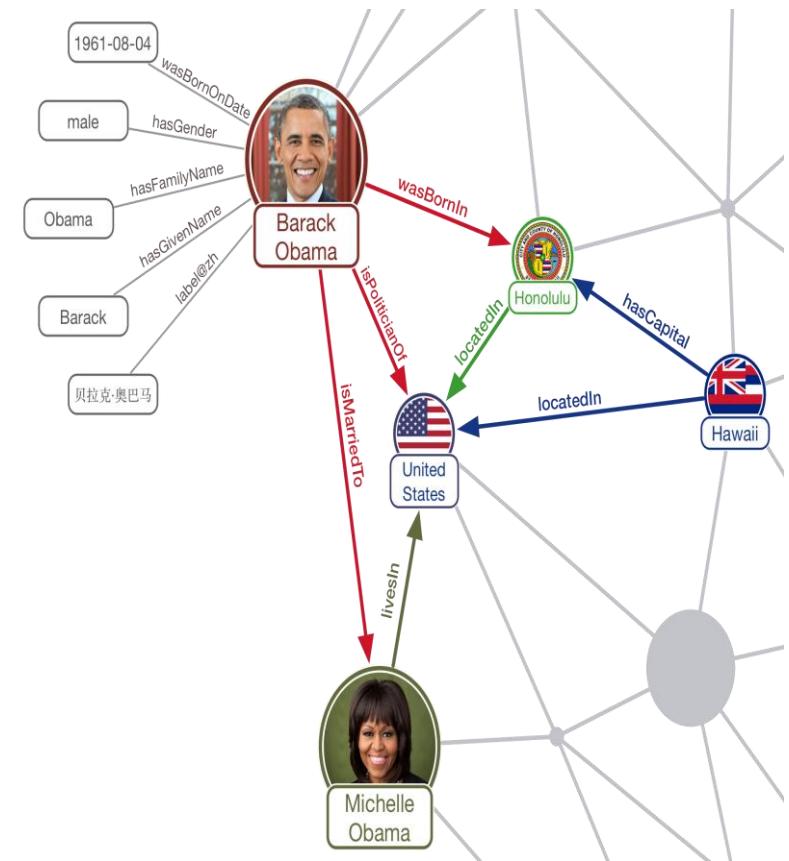
Why we need Knowledge Graphs

- Motivation and Solution
- Touristic Use Cases
- Energy Use Cases
- Further Verticals

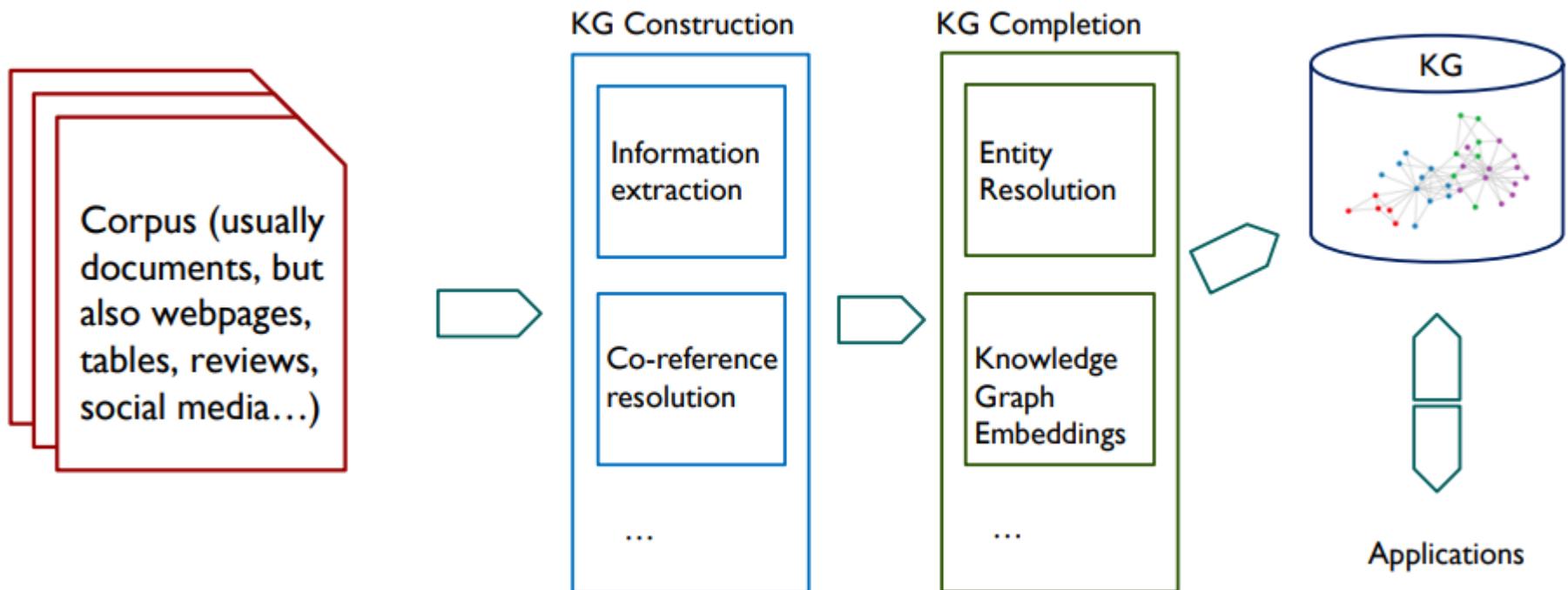
KG Example

Set of triples, where each triple (h, r, t) represents a **relationship r** between **head entity h** and **tail entity t**

(Barack Obama, wasBornOnDate, 1961-08-04),
 (Barack Obama, hasGender, male),
 ...
 (Hawaii, hasCapital, Honolulu),
 ...
 (Michelle Obama, livesIn, United States)



KG Workflow



Information Extraction

Isaac Newton

From Wikipedia, the free encyclopedia.

<Person>

Sir Isaac Newton (25 December 1642 – 20

March 1727 by the Julian calendar in use in

England at the time; or 4 January 1643 – 31

March 1727 by the Gregorian calendar) was an

English physicist, mathematician, astronomer,
philosopher, and alchemist; who wrote the

Philosophiae Naturalis Principia Mathematica

(published 5 July 1687)¹, where he described

universal gravitation and, via his laws of motion,

laid the groundwork for classical mechanics

Newton also shares credit with Gottfried Wilhelm

Leibniz for the development of differential calculus.

However, their work was not a collaboration; they both discovered

calculus separately but nearly contemporaneously.

<TimePeriod>

<Scientist>

<Scientist>

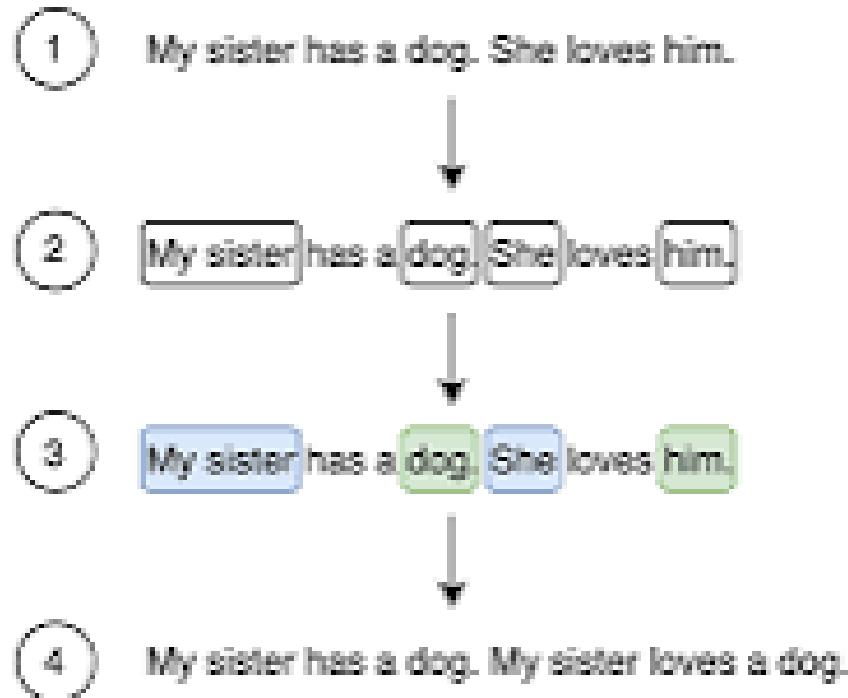


Sir Isaac Newton
in Kneller's <Painter>
portrait of 1689.

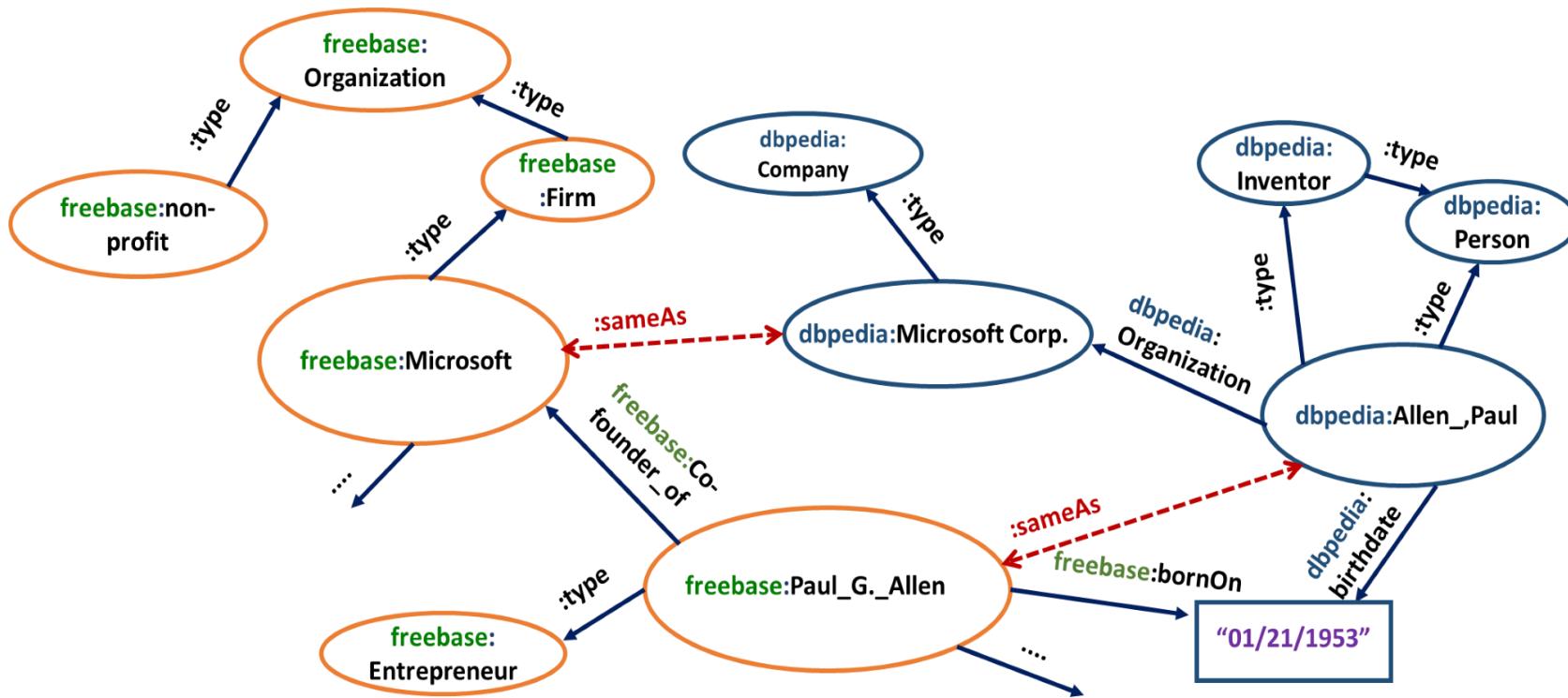
<Publication>

<Person>

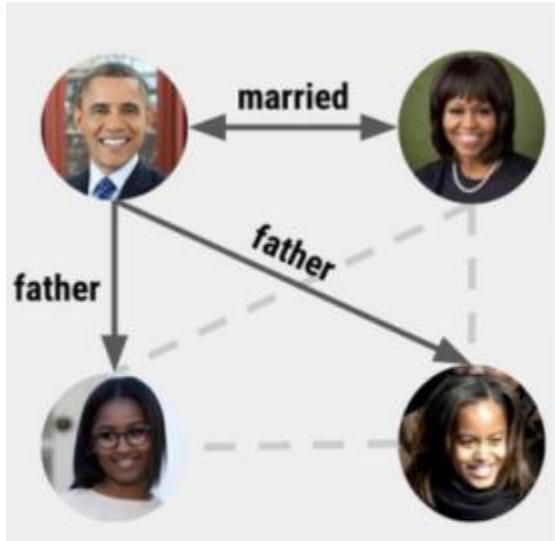
Coreference Resolution

- 1 My sister has a dog. She loves him.
 - 2 My sister has a dog. She loves him.
 - 3 My sister has a dog. She loves him.
 - 4 My sister has a dog. My sister loves a dog.
- 

KG Example

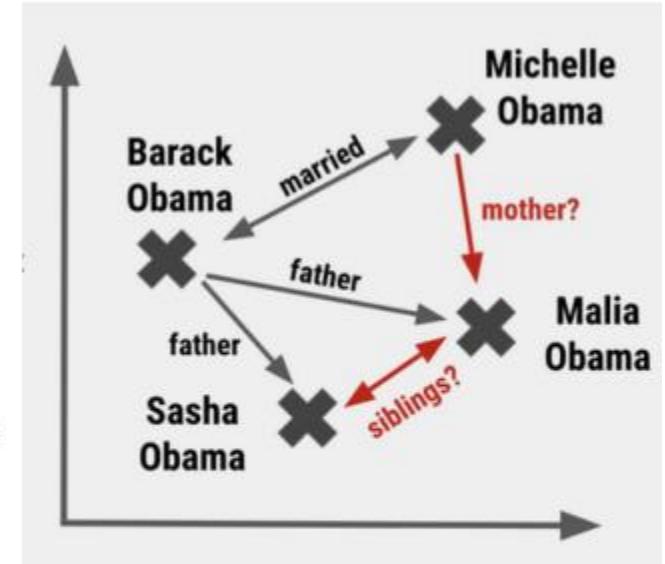


Knowledge Graph Embedding

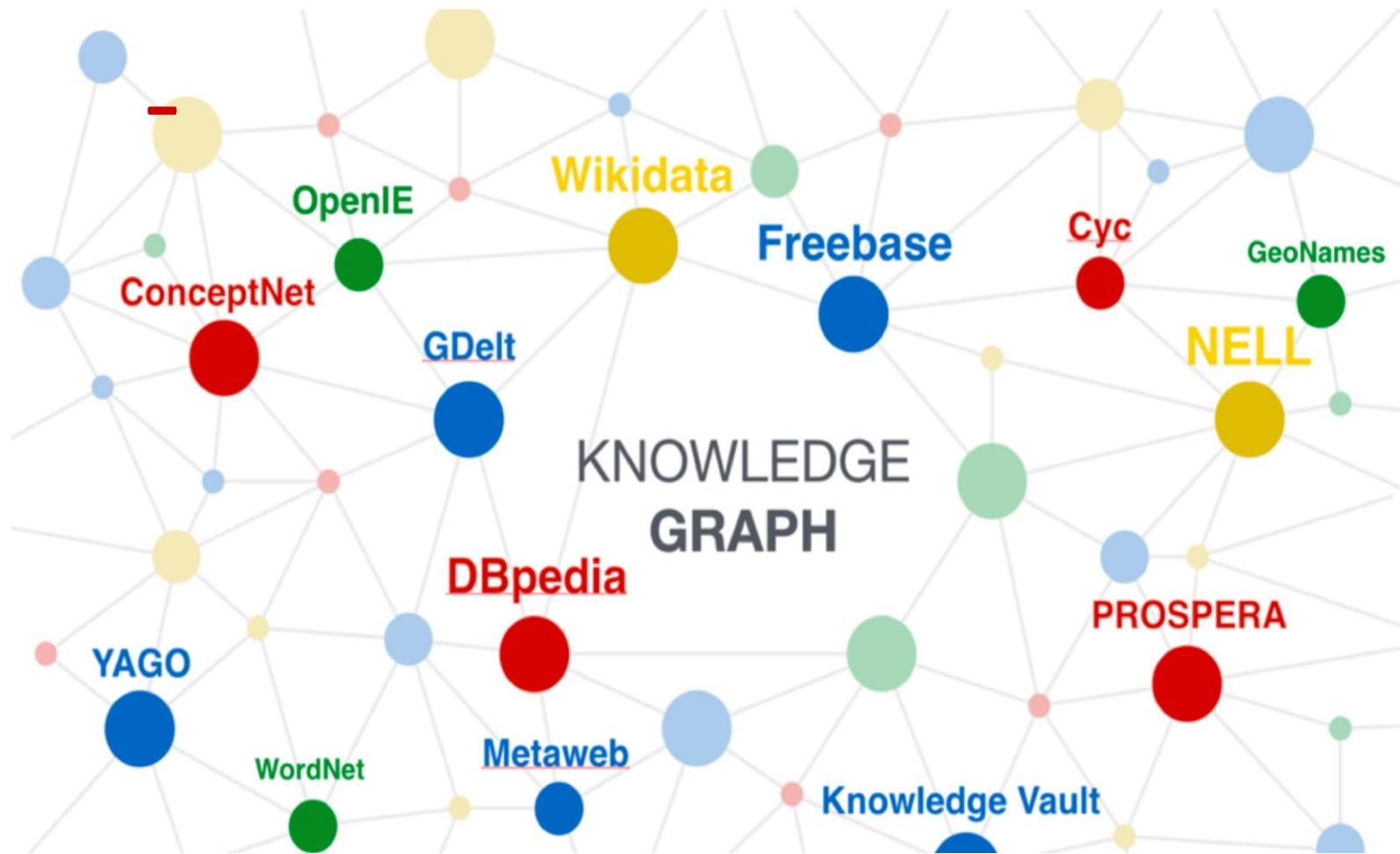


Knowledge graph embeddings:

- TransE, H...
- Neural tensor networks
- Graph convolutional networks (or their variants)
- Matrix factorization
- ...

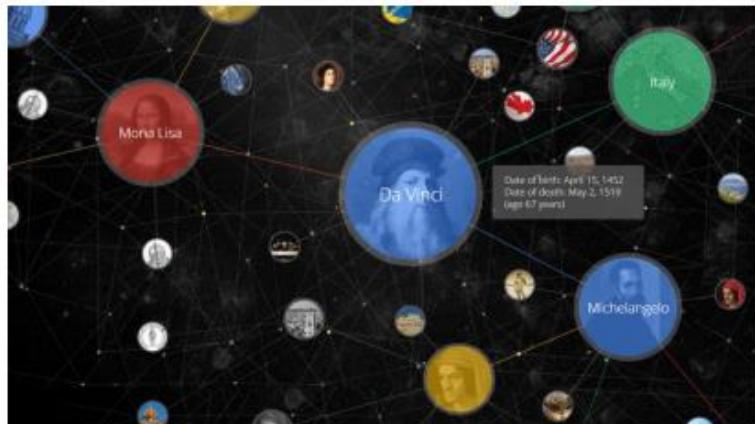


Open source KGs



Google Knowledge Graph

Knowledge Graph Example 1: Leonardo da Vinci



About 26,700,000 results (0.67 seconds)

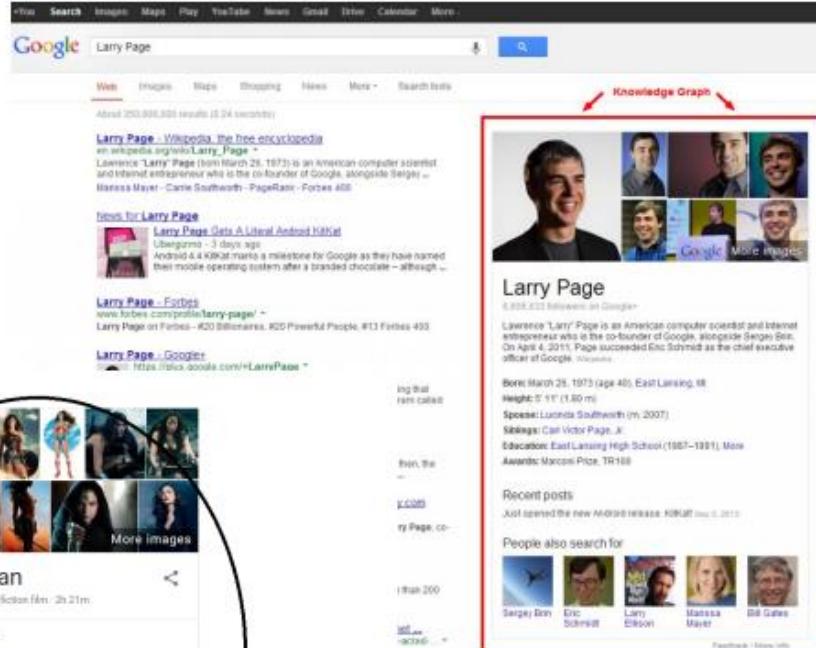
Wonder Woman (2017) - IMDb
<https://www.imdb.com/title/tt6451270/> ★★★★☆ Rating: 7.6/10 - 360,588 votes
 When a pilot crashes and finds herself in the outside world, Diana, an Amazonian warrior in training, learns how to fight and discover her full powers and true destiny.
 Full Cast & Crew - Chris Pine - Trivia - Parents Guide

Wonder Woman (2017 film) - Wikipedia
[https://en.wikipedia.org/wiki/Wonder_Woman_\(2017_film\)](https://en.wikipedia.org/wiki/Wonder_Woman_(2017_film)) Wonder Woman is a 2017 American superhero film based on the DC Comics character of the same name, developed by Warner Bros. Pictures. It is the fourth installment in the DC Extended Universe (DCU). The film is directed by Patty Jenkins, with a screenplay by Allan Heinberg, from a story by Jenkins, Zack Snyder, ...
 Gal Gadot, Patty Jenkins, Elena Anaya, Doctor Pesci

Top stories

- Oscars voting ends today. Will Wonder Woman finally break the anti-superhero streak? [Washington Post](#)
- Fashion War: Wonder Woman Gal Gadot Inflames Lebanese with Dress Design [Breitbart](#)
- Gal Gadot Diet and Fitness Routine | POPSTAR Fitness Australia [POPSUGAR Australia](#)

Knowledge Graph Example 2: Larry Page



Google Larry Page

Web Images Maps Shopping News More Search tools

About 250,000,000 results (0.24 seconds)

Larry Page - Wikipedia, the free encyclopedia
https://en.wikipedia.org/w/index.php?title=Larry_Page Lawrence "Larry" Page (born March 26, 1973) is an American computer scientist and Internet entrepreneur who is the co-founder of Google, alongside Sergey Brin. ...
 Marissa Mayer - Carrie Southworth - PageRank - Forbes 400

News for Larry Page

Larry Page Gets A Literal Android Head
<https://www.ubergizmo.com/poll/larry-page/> -
 UberGizmo - 3 days ago
 Android 4.4 KitKat marks a milestone for Google as they have named their mobile operating system after a braided chocolate - although ...

Larry Page - Forbes
<https://www.forbes.com/poll/larry-page/> -
 Larry Page is on Forbes' 100 Billionaires, #20 Powerful People, #13 Forbes 400.

Larry Page - Google
<https://plus.google.com/+LarryPage> -

Recent posts
 Just learned the new movie release. [#WonderWoman](#) (Dec 8, 2017)
 People also search for
 Sergey Brin, Eric Schmidt, Larry Ellison, Marissa Mayer, Bill Gates

Feedback | Privacy info

KG Application

places to visit san jose

All Maps Images News Shopping More Settings Tools

About 77,800,000 results (1.02 seconds)

Top things to do in San Jose

- Winchester Mystery House
Quirky mansion with odd design details
- The Tech Interactive
Interactive displays & an IMAX theater
- Mission Peak
Mountain with an iconic summit pole
- Happy Hollow Park & Zoo
Animals, activities & conservation focus

[More things to do](#)

Things To Do & Places To See in San Jose | Visit San Jose

<https://www.sanjose.org/things-to-do> ▾
Vibrant, diverse and accessible, San Jose shines with 300+ days of sunshine that assures indoor and outdoor activities for every inclination. Whether it's nightlife ...
Attractions and Amusement - Kids and Family · San Jose Wineries · Nightlife

Complete List of San Jose Attractions | Visit San Jose

<https://www.sanjose.org/things-to-do/attractions-and-amusement> ▾
Find a complete list of San Jose attractions and things to do - perfect for tourists and ... Browse the best places to see in San Jose and the surrounding Bay Area.

25 Best Things to Do in San Jose (CA) - The Crazy Tourist

<https://www.thecrazytourist.com> › ... › United States › California (CA) ▾
25 Best Things to Do in San Jose (CA): Winchester Mystery House: Flickr. Tech Museum Of Innovation: Flickr. Children's Discovery Museum: Flickr. Rosicrucian Egyptian Museum: Flickr. San Jose Heritage Rose Garden: wikimedia. Basilica Of St. Joseph: Flickr. Alum Rock Park: Flickr. Happy Hollow Zoo Entrance: Flickr.

THE 15 BEST Things to Do in San Jose - 2019 (with Photos ...

https://www.tripadvisor.com/Attractions-g33020-Activities-San_Jose_Calif... ▾
... attractions. Find what to do today, this weekend, or in August. We have reviews of the best places to see in San Jose. Visit top-rated & must-see attractions.
Free Entry (36) · Things to Do in San Jose · Museums in San Jose · Santana Row

San Jose
City in California

San Jose is a large city surrounded by rolling hills in Silicon Valley, a major technology hub in California's Bay Area. Architectural landmarks, from the 1883 Italianate-style Oddfellows building to Spanish Colonial Revival structures, make up the downtown historic district. The downtown area is also home to the Tech Museum of Innovation, devoted to the exploration of science and technology.

Weather: 64°F (18°C), Wind NW at 10 mph (16 km/h), 85% Humid
Population: 1.035 million (2017)

Plan a trip

- San Jose travel guide
- 3-star hotel averaging \$206
- 1 h 5 min flight, from \$97

Did you know? San Jose, California has the largest Vietnamese-American population (106,992) among all U.S. cities. [wikipedia.org](https://en.wikipedia.org)

People also search for

- San Francisco
- Santa Clara County
- California
- San Diego
- San Francisco Bay Area

[View 15+ m](#)

Knowledge panel

Recognition of user intent

Recommendations

Exploration suggestions

KG Motivation and Solution



KG Motivation and Solution

- A dialogue with an Amazon Echo or Google Home quite often ends in “Sorry, I do not know.” due to the lack of domain-specific knowledge these devices have
- To support the chatbot and voice assistant, need to design, implement, and deploy a knowledge-centered solution

KG Motivation and Solution

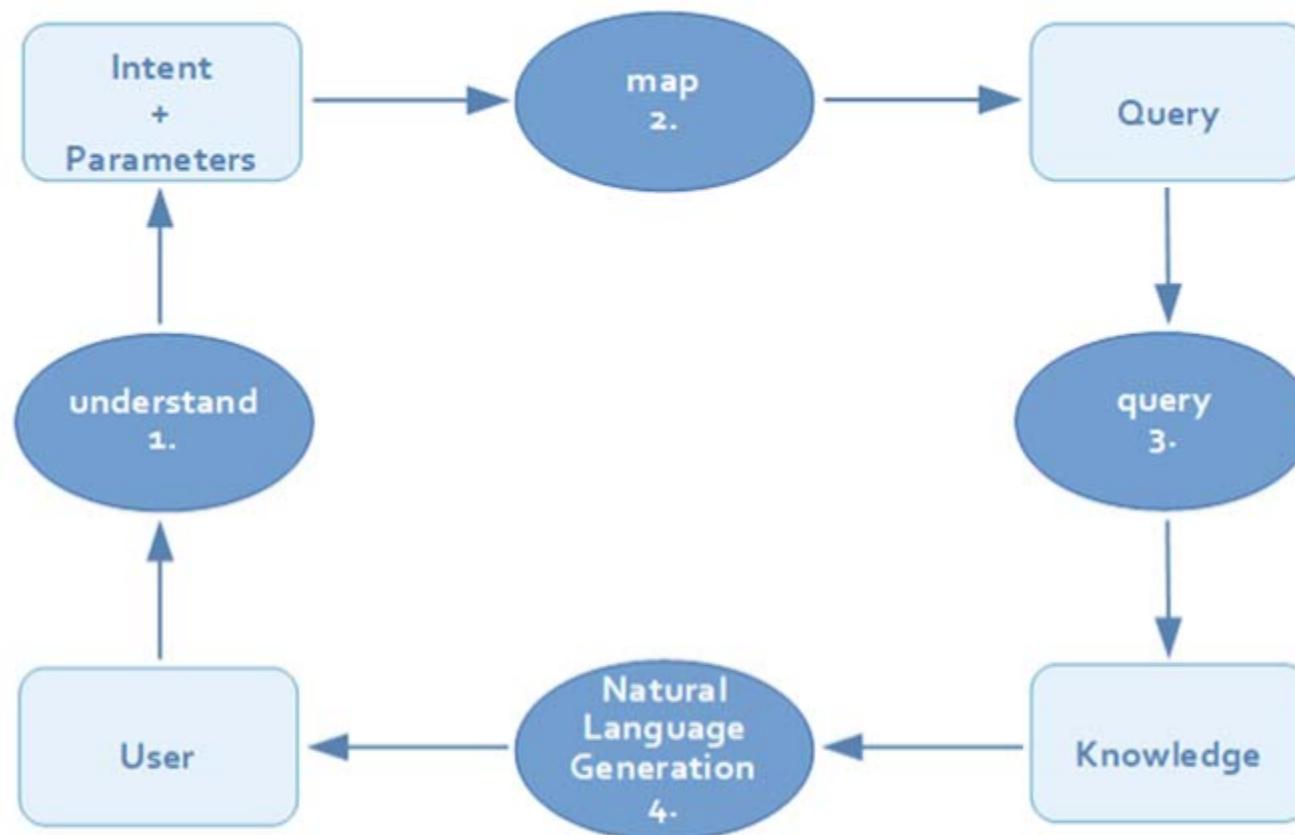


Fig. 4.2. The inner process of a Knowledge-centered Chatbots and Voice Assistants²²⁰

KG Motivation and Solution

- Natural language input of a user, in written or spoken form, undergoes a natural language understanding step (understand 1.), in which the user intent, together with parameters, are identified.
- The intent needs then to be resolved to an action that typically translates in a set of queries (map 2.)
- This is then be executed (query 3.) against large volumes of heterogeneous, distributed, dynamic, and potentially (i.e., nearly for sure) inconsistent statements in order to identify the relevant knowledge parts
- Generate the user answer in natural language
- Chatbots and voice assistants are fed with knowledge from the Knowledge Graph in order to better answer human users' natural language queries and engage in meaningful dialogues with them

Knowledge-centered Chatbots and Voice Assistants



Conversational Platform

Platform Features

- ✓ Multichannel (voice assistant, chat, phone)
- ✓ Conversation modules and templates
- ✓ Content/Data plugins
- ✓ Conversation management per channel
- ✓ Connector for external data and services
- ✓ Manage conversation content
- ✓ Analytics of conversations

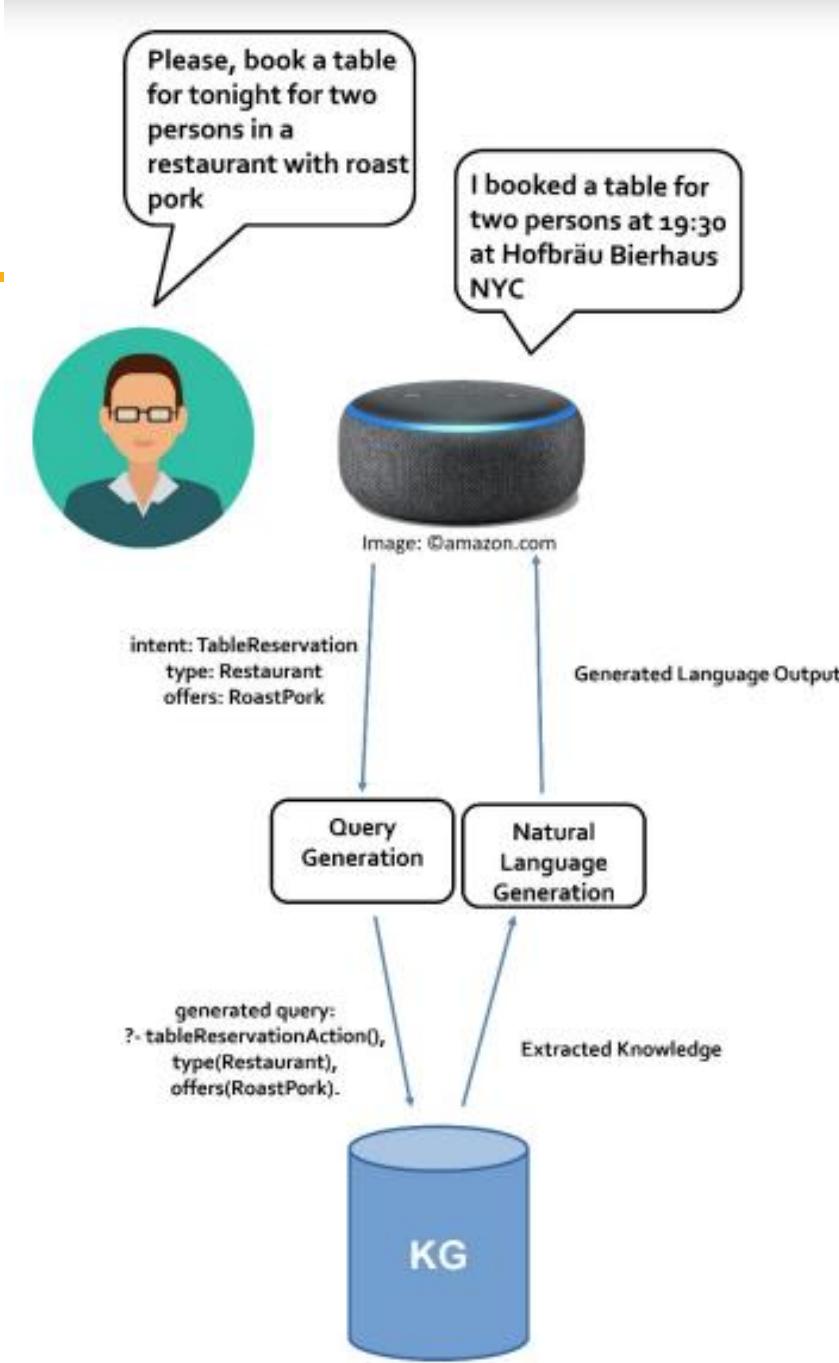
Knowledge Graph Platform

Platform Features

- ✓ Data Integration
- ✓ Data Life Cycle
- ✓ Data Modeling
- ✓ Entities
- ✓ Relationships
- ✓ Semantics

KG Motivation and Solution

- Algorithms that run on top of Knowledge Graphs include inference, recommendations, machine learning, text understanding etc
 - Inference algorithms, of various complexity, e.g., rule based reasoning, OWL/RDFS reasoning or combinations of these approaches can be used to infer and enrich the Knowledge Graph
 - use the structure and semantics of data to generate knowledge for chatbots and voice assistants in the form of entities and actions and combine this with personalization
 - leads to more relevant content, additional interesting facts, in the short-term better recommendations for a user and better results for content providers as well as, over time, smarter chatbot system
 - derive from the Knowledge Graph training data for entity recognition (e.g., Telfs is a village) and provide (semi-)automatically generated intents and example questions
-



- Intent Table Reservation, for entities of type Restaurant can be generated.
- Restaurants, and in general, organizations can be connected in the Knowledge Graphs to other entities of type Offer. (e.g., Roast Pork offers)
- Generate the rules that restrict the view/access to the Knowledge Graph depending on the use cases.
- Such rules, together with the intent and parameters extracted by the NLU, are used to generate the queries to be executed against the Knowledge Graph
- Knowledge Graph can be used to generate templates for the answers, the textual answers or follow up questions to run the dialogues.

Tourist Chatbot Use Case

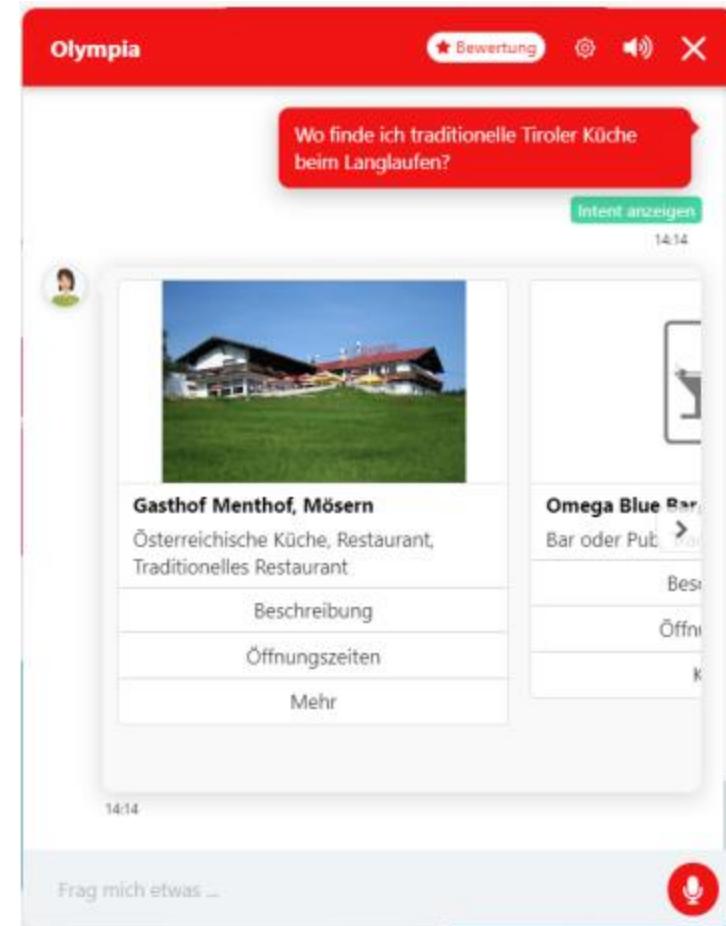
- Planning phase hotels and Destination Management Organizations (DMOs) can provide information about the hotel and the region, the surroundings, and weather condition
 - Booking phase, from booking the hotel and transport to buying connected services, e.g., ski tickets, all becomes much simpler and efficient by using natural language.
 - Experiences phase, Chatbots and Voice Assistants can also announce special offers or events.
 - All requested information and processes are available 24/7/365 and instantly
-

Tourist Chatbot Use Case

- Need structures that encode the knowledge about the tourism domain, in terms of entities and relations between them, in a machine processable form.
- Knowledge Graphs are such structure providing the technical means to integrate various heterogeneous touristic information sources
- Integrates multiple sources of content, data, and services from various providers, both
 - Closed sources: feratel, General Solutions, intermaps, Outdooractive etc
 - Open sources: DBpedia, GeoNames, OpenStreetMap, and Wikidata

Tourist Chatbot Use Case

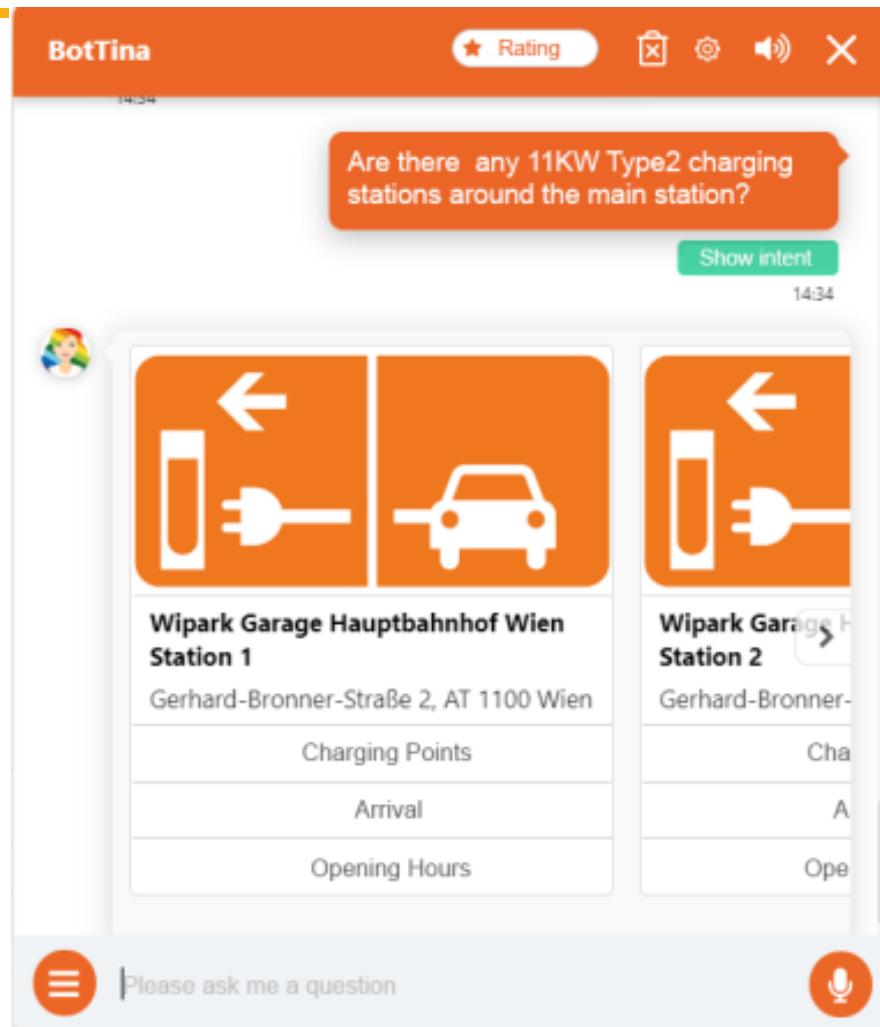
"Where can I have a traditional Tyrolean food when going cross country skiing?"



Tourist Chatbot Use Case

- Need to integrate data from multiple heterogeneous static and dynamic sources for which we need to track provenance and maintain one shared evolving schema
- Ensured a level of quality of the knowledge.
- Optimize conversational interfaces based on Knowledge Graphs by providing a rich intent and entity management (e.g., automated NLU training), question answering over the Knowledge Graph
- Supporting advanced dialogues such as guiding a user through actions or recommendations or follow up conversations

Energy Use Cases



BotTina

Rating

Are there any 11KW Type2 charging stations around the main station?

Show intent

14:34

14:34

Wipark Garage Hauptbahnhof Wien Station 1

Gerhard-Bronner-Straße 2, AT 1100 Wien

Charging Points

Arrival

Opening Hours

Wipark Garage H Station 2

Gerhard-Bronner-

Cha

A

Ope

Please ask me a question

Wien Energy Knowledge Graph

- Includes information about the electric vehicle charging stations in the city of Vienna.
- Extended the schema.org vocabulary with additional types and properties.
- Can model information such as the different types of plugs available at an EV charging station, its voltage and power, its address and geo-coordinates, as well as accepted payment methods.
- KG populated with data received from Wien Energie EV charging station API.
- Static data and dynamic data considered. Ex no. of available plugs, waiting times and offers
- Using the Knowledge Graph, the chatbot can answer questions such as: "Are there any 11KW Type 2 charging stations around the main station?"

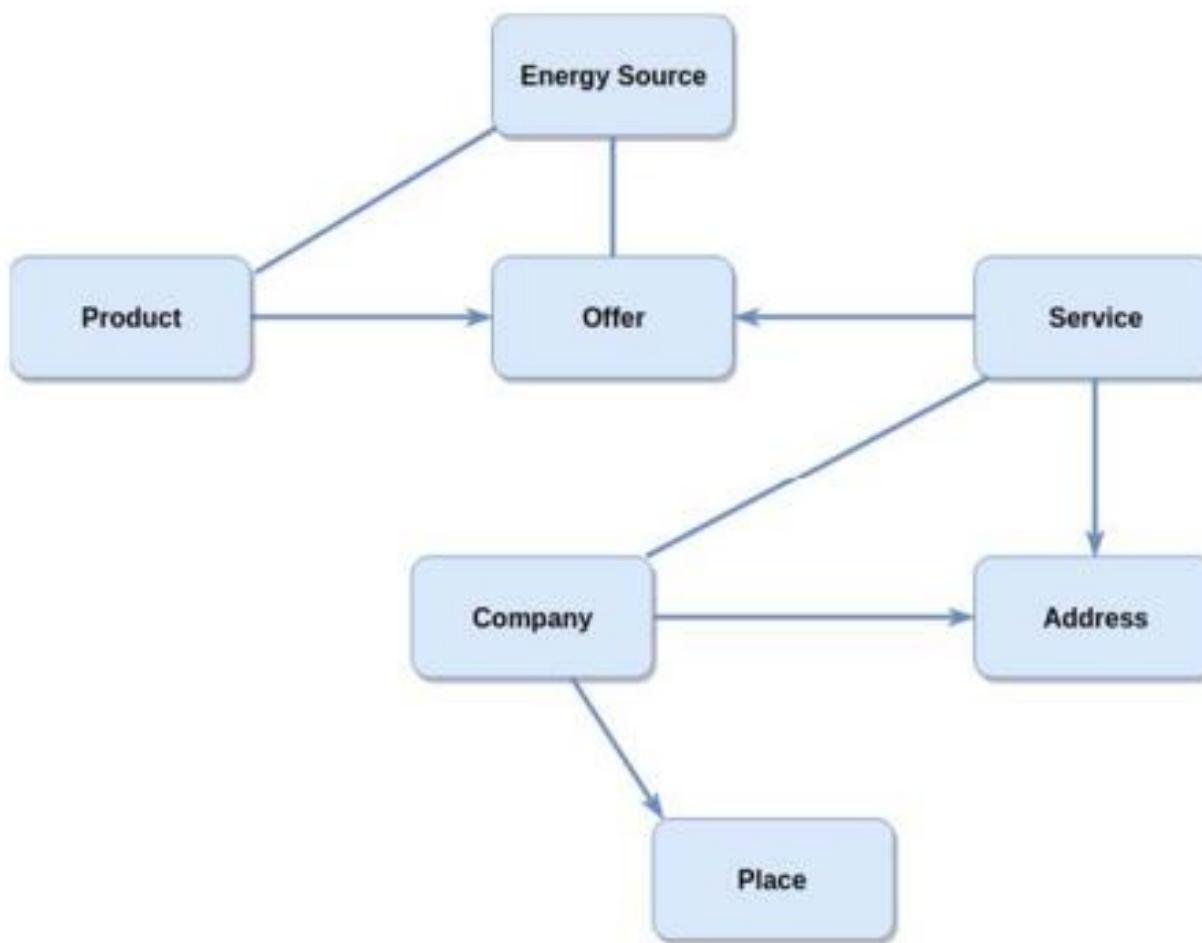
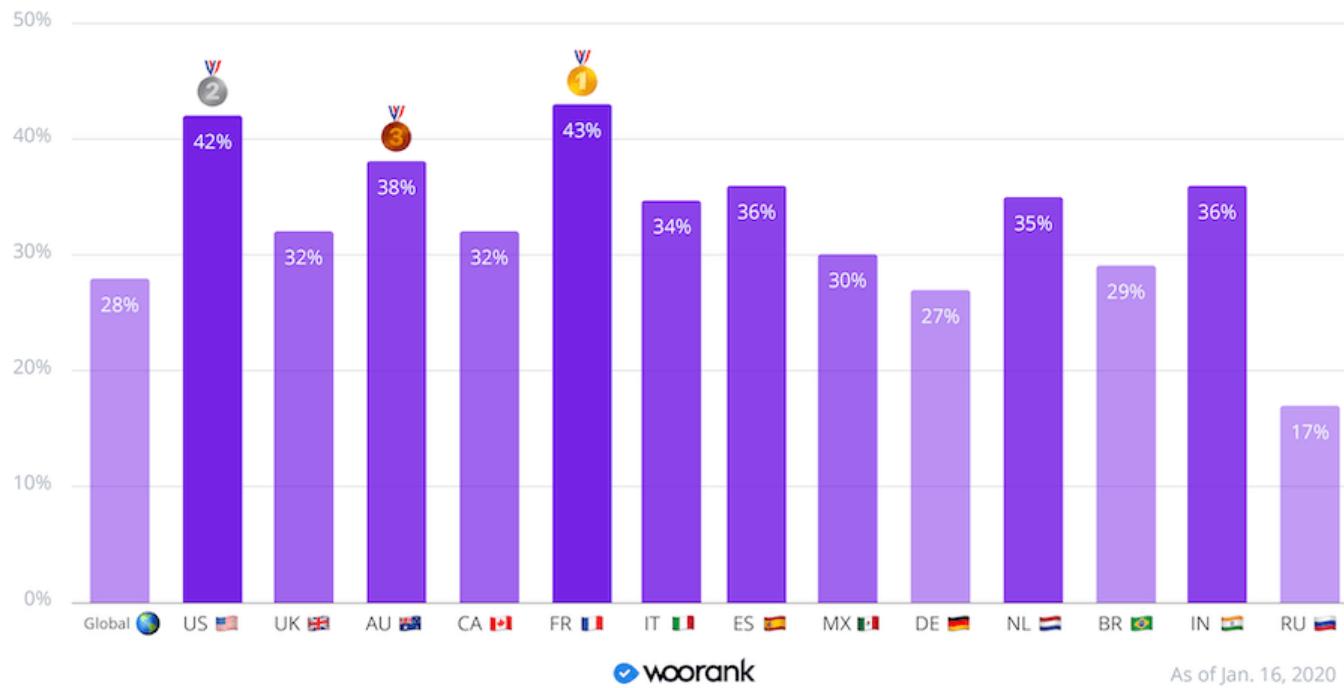


Fig. 4.8. Energienetze Steiermark domain model top-level types

Schema.org use by country

Percentage of websites reviewed by WooRank that use Schema.org structured data



Other Verticals: Education

The University of Innsbruck pilot

- Constructing a Knowledge Graph about study programs offered
- Used to improve conversations related to the different study programs, subjects, and examinations.
- Based on the entities and relations modeled in the Knowledge Graph, dynamic intents are created to query the content from the Knowledge Graph
- Answer questions such as “Which study programs are offered by the Faculty of Biology?”, or ‘How many semesters do I need to study for the Master’s Programme Business Law?’

Other Verticals: Finance

Wiener Börse pilot

- focuses on modeling knowledge about companies, equities, bonds, indexes available on the Vienna Stock Exchange
- Accessing data in real time
- High-frequency data are essential to enable the Chatbot and Intelligent Personal Assistant to provide accurate, up-to-date information
- Chatbot answers questions such as "What is the stock price of X?", "Which stocks have lost the most?"

Other Verticals: Retail

- Radically change the e-marketing and ecommerce processes.
 - Customers will be able to search for products that fit their needs and then may buy these products, by directly engaging in natural dialogues with the retail Chatbots and Intelligent Personal Assistants.
 - Access to clothing and footwear products, models and integrates information about product catalogs, product stocks, shops, and connected services into a retail Knowledge Graph
 - Answering complex questions about products or e-commerce processes
-

KG Growth

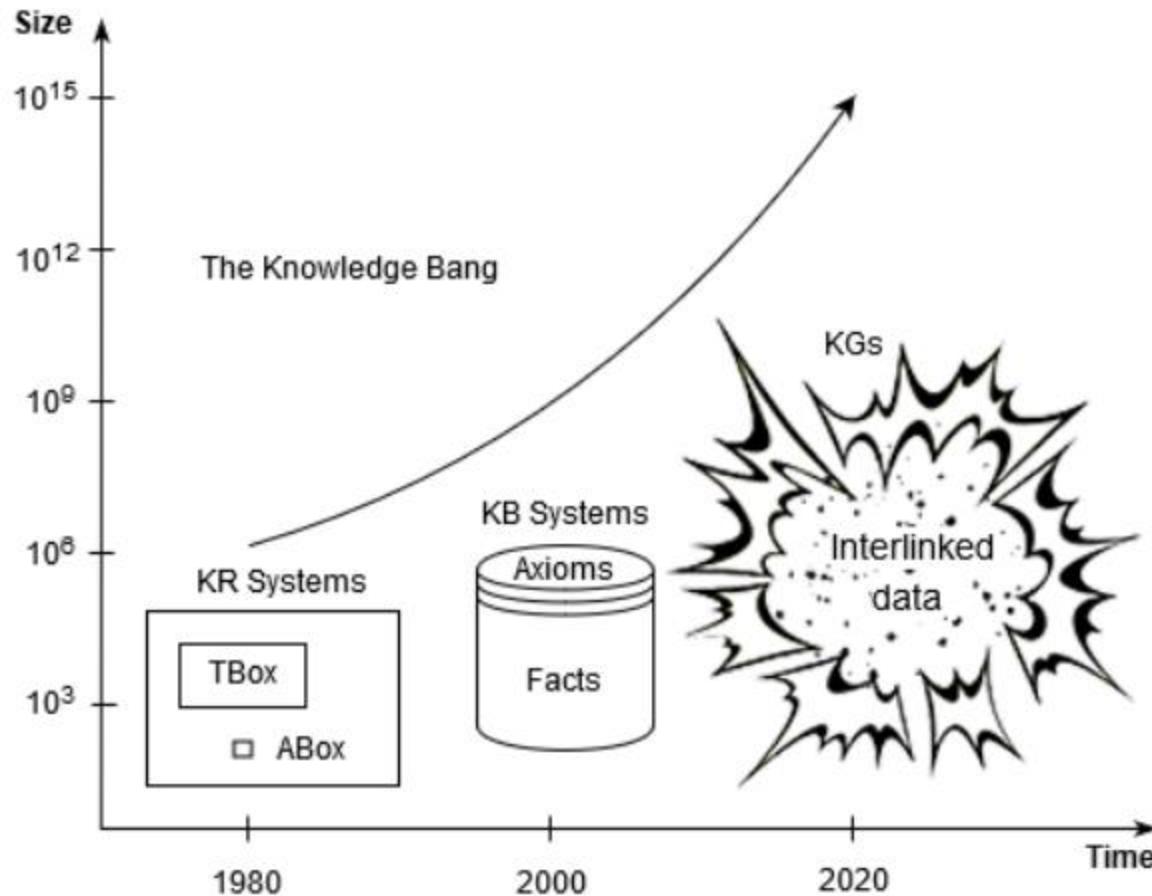


Fig. 5.1 The inflationary growth of the Knowledge Universe

New KG Systems

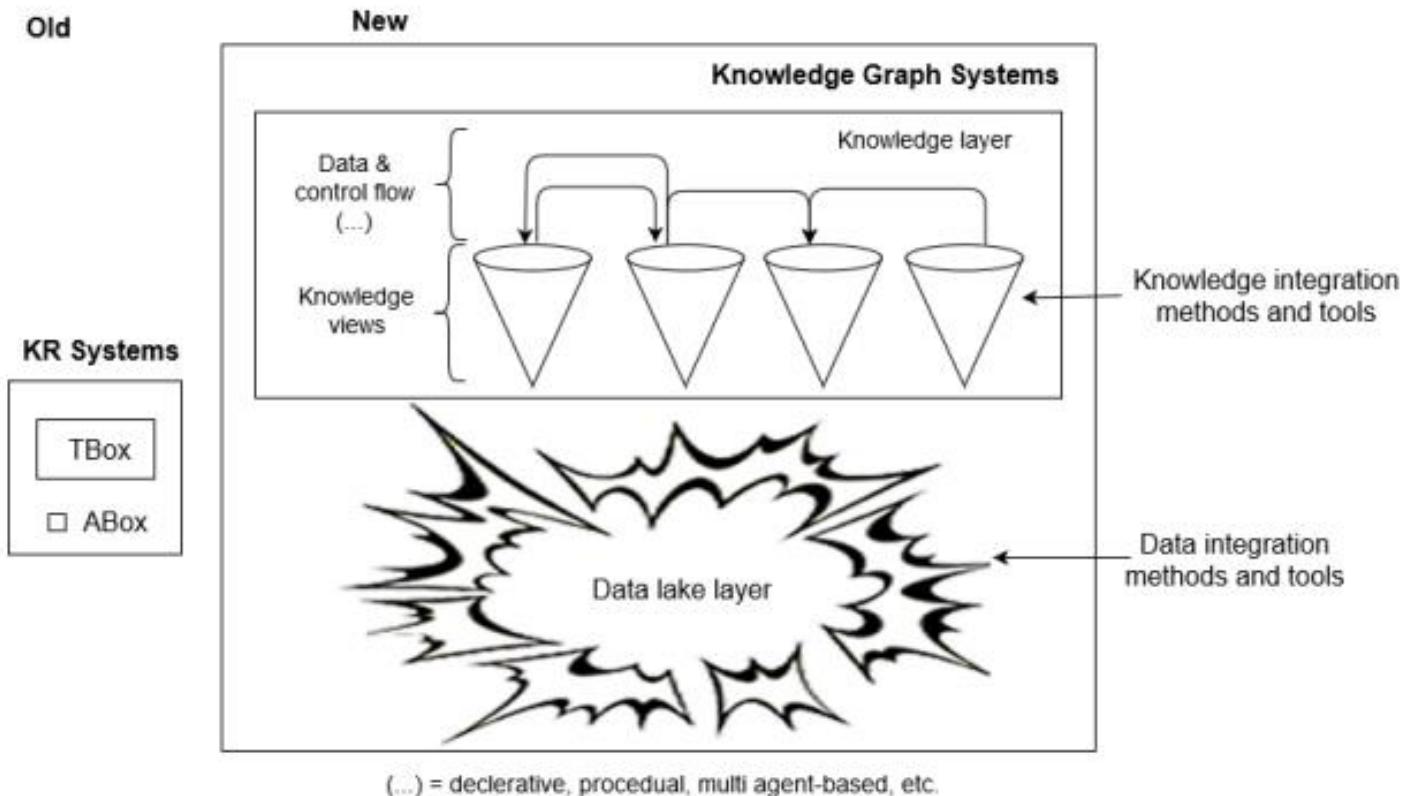


Fig. 5.2 Knowledge Graph System Architecture.

Open Information Extraction

ID	Document
1	Your dry cleaner set out from eastern Queens on foot Tuesday morning and now somewhere near Maspeth.
2	Recently, North Korea has begun to allow tourists, including Americans, ..., and South Korean tourists have been able to go to Kaesong on a limited basis.
...	...

Entity 1	Relation Phrase	Entity 2	Human Evaluation
your dry cleaner	<i>set out from</i>	eastern Queens	✓
your dry cleaner	<i>set out from_on</i>	foot	✓
your dry cleaner	<i>is near</i>	Maspeth	✓
North Korea	<i>has begun to allow</i>	tourist	✓
South Korean tourists	<i>to go to</i>	Kaesong	✓
...	
Queens	<i>on</i>	foot	✗
Kaesong	<i>on</i>	a limited basis	✗

Coreference Resolution, Entity Linking...

"I had no idea I was getting in so deep," says Mr. Kaye, who founded Justin in 1982. Mr. Kaye had sold Capetronic Inc., a Taiwan electronics Maker, and retired, only to find he was bored. With Justin, he began selling toys and electronics made mostly in Hong Kong, beginning with Mickey Mouse radios. The company has grown -- to about 40 employees, from four initially, Mr. Kaye says. Justin has been profitable since 1986, adds the official, who shares [his] office... (nw/wsj/2418)

The Northern Lights, also called Aurora Borealis, are one of the most spectacular shows on this earth and can frequently be seen in Iceland from September through March on clear and crisp nights.



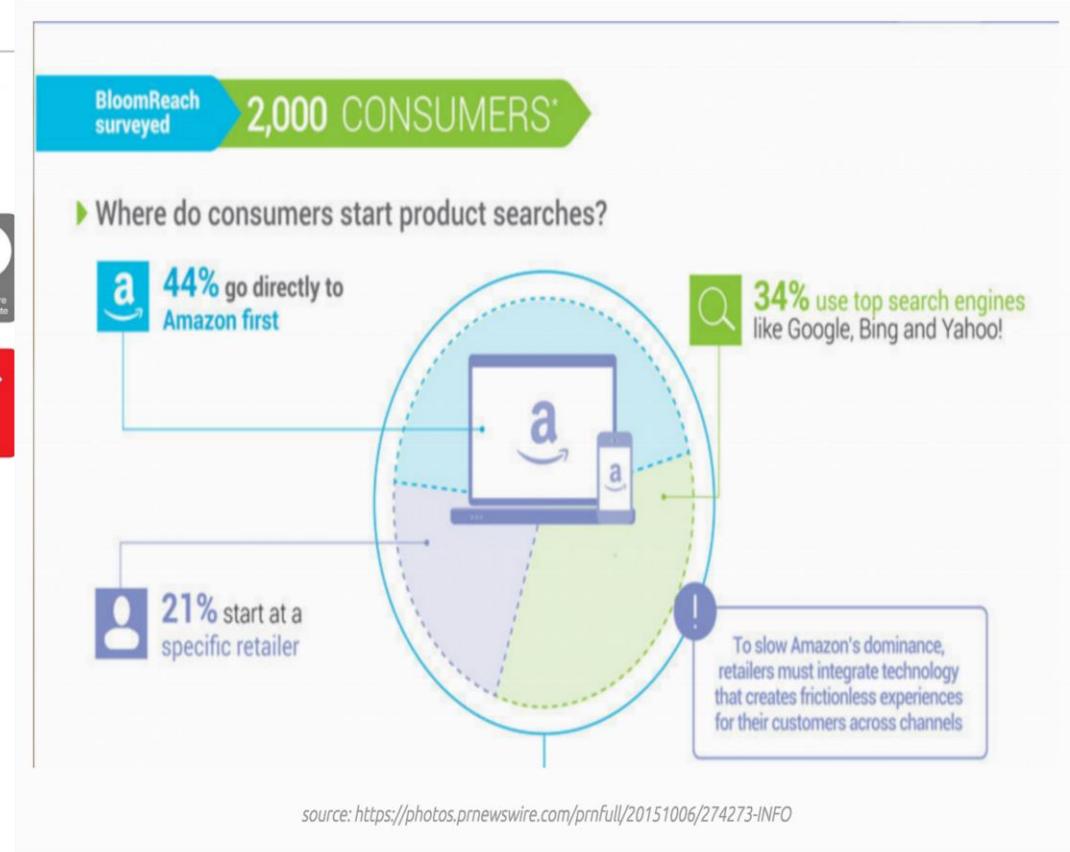
-----Finding entities belonging to category : place name
[0..1) location : Iceland



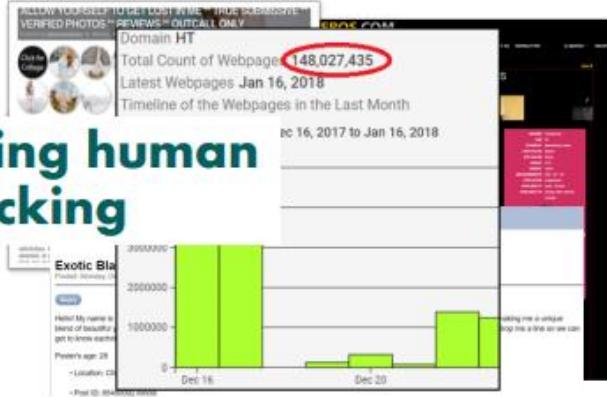
Iceland, California
From Wikipedia, the free encyclopedia
Iceland
Nevad
It had
about
Iceland
area. 1
the sout
Refe
Hydr
1
2
Iceland Lake
From Wikipedia, the free encyclopedia
Iceland
the sout
Refe
Hydr
1
2
Iceland
From Wikipedia, the free encyclopedia
This article is about the country. For other uses, see Iceland (disambiguation).

Domain Specific Search

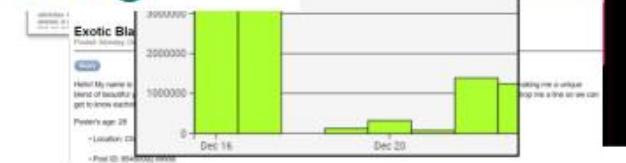
The Massive YouTube Ecosystem



Emerging applications of KG



Fighting human trafficking



Penny Stock Fraud Nets Millions

Scheme Mastermind Among Those Sentenced to Prison



By July 11, 2004

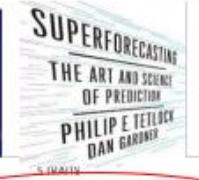
Immigration is increasingly taking legal action against individuals and companies that file. In one of its recent cases, involving Ives Health Co., the SEC reported a final her president, M. Keith Ives, for disseminating misleading information on the Internet.

State investors a total of \$1.25 million for, among other things, falsely claiming the company developed.

Stopping Penny Stock Fraud

Defined by the SEC as stocks that sell below \$5 a share, penny stocks have always been considered speculative and easily manipulated. But stock market experts, seeing an increase in penny stock promotion online, say investors should be wary of

Predicting cyberattacks



Accurate geopolitical forecasting



- 1. Superforecasters begin by gathering as much information possible.
- 2. Superforecasters make and develop the habit of thinking terms of probabilities when exploring the likelihood of specific events.
- 3. Superforecasters improve when individuals work in teams.
- 4. Superforecasters ensure that they are regularly keeping score of their projections.
- 5. The most successful forecasters are willing to admit error and quickly change course on their projections.

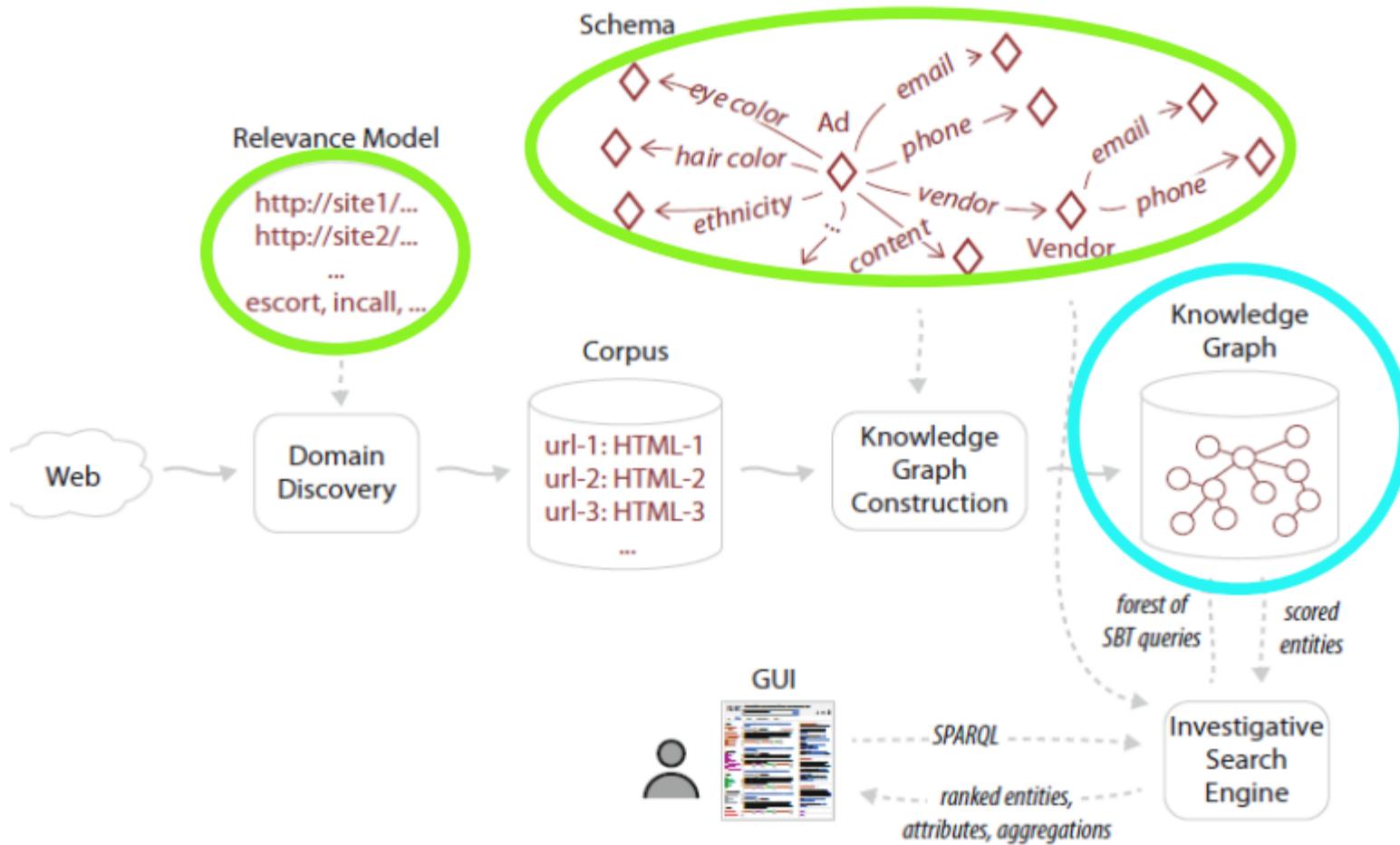
Research question

General Search
DSS

Google Knowledge Graph
Domain-Specific Knowledge Graphs

How do we construct domain specific knowledge graphs over web data for powerful DSS applications?

Knowledge graph for DSS



Amazon Product Graph

❑ Mission: To answer any question about products and related knowledge in the world



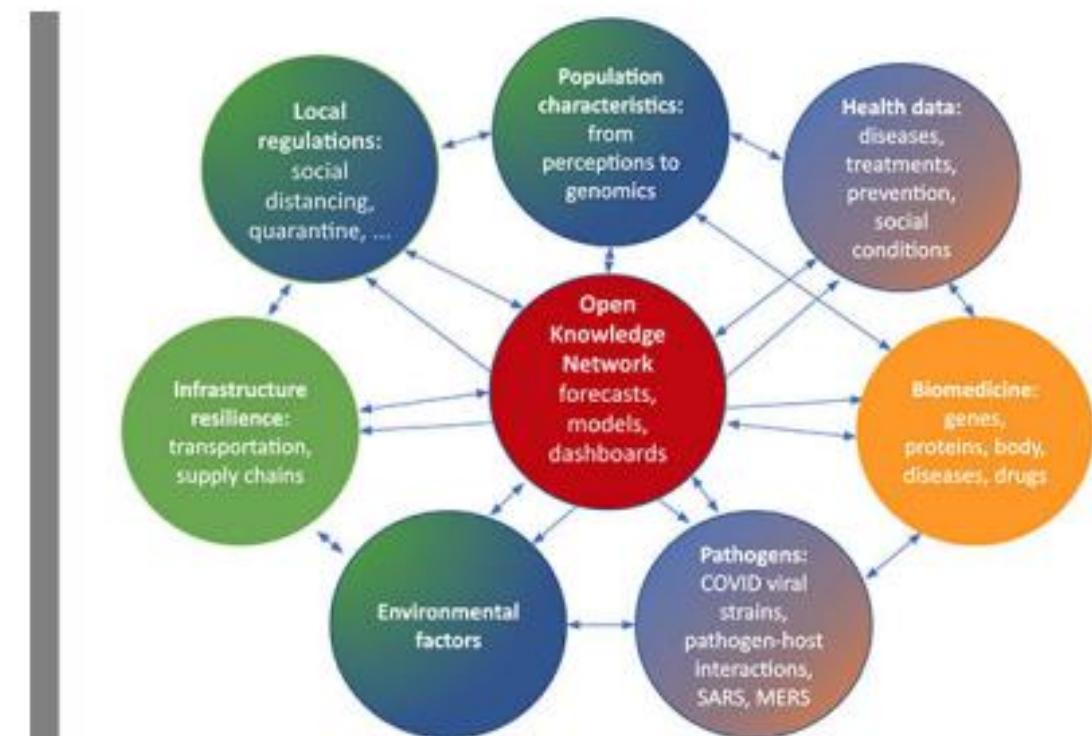
Source: Dong, Luna. Building a Broad Knowledge Graph for Products. Keynote at ICDE. 2019

June 01, 2020 | By Jan Zverina

Fighting COVID-19 with Knowledge Graphs

National Science Foundation awards funding for a semantic integration platform

"The project will be based on our knowledge graph prototype linking information about pathogens, health data, and environmental indicators and enabling cross-domain inferencing," said Peter Rose, director of SDSC's Structural Bioinformatics Laboratory and principal investigator (PI) for the project, called 'COVID-19-Net: Integrating Health, Pathogen and Environmental Data into a Knowledge Graph for Case Tracking, Analysis, and Forecasting.' "Such a graph lets researchers trace the spread of the coronavirus in different geographic conditions, focusing on specific virus strains and transmissions."

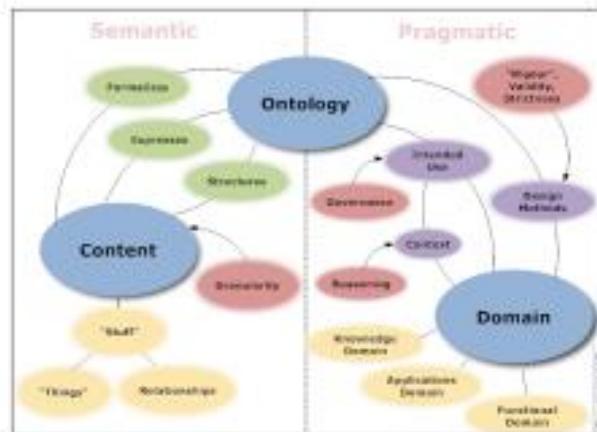


What is (isn't) domain?



IT'S IN THE SYLLABUS

This message brought to you by every instructor that ever lived.
WWW.PICCOMICS.COM
'Read Higher and Deeper' by Jorge Chávez



Some dictionary definitions

(Merriam Webster) A sphere of **knowledge, influence or activity**

(Oxford) A **specified** sphere of activity or knowledge

Specifying the sphere

Rules

Scope (e.g., the legal system)

Syllabi (for classrooms)

Examples

How do domain experts specify the sphere?

Examples

Ontology

Knowledge Graph in Personal Assistant



Taylor Swift > Songs



Love Story
Fearless · 2008



Look What You Made Me...
Reputation · 2017



Shake It Off
1989 · 2014



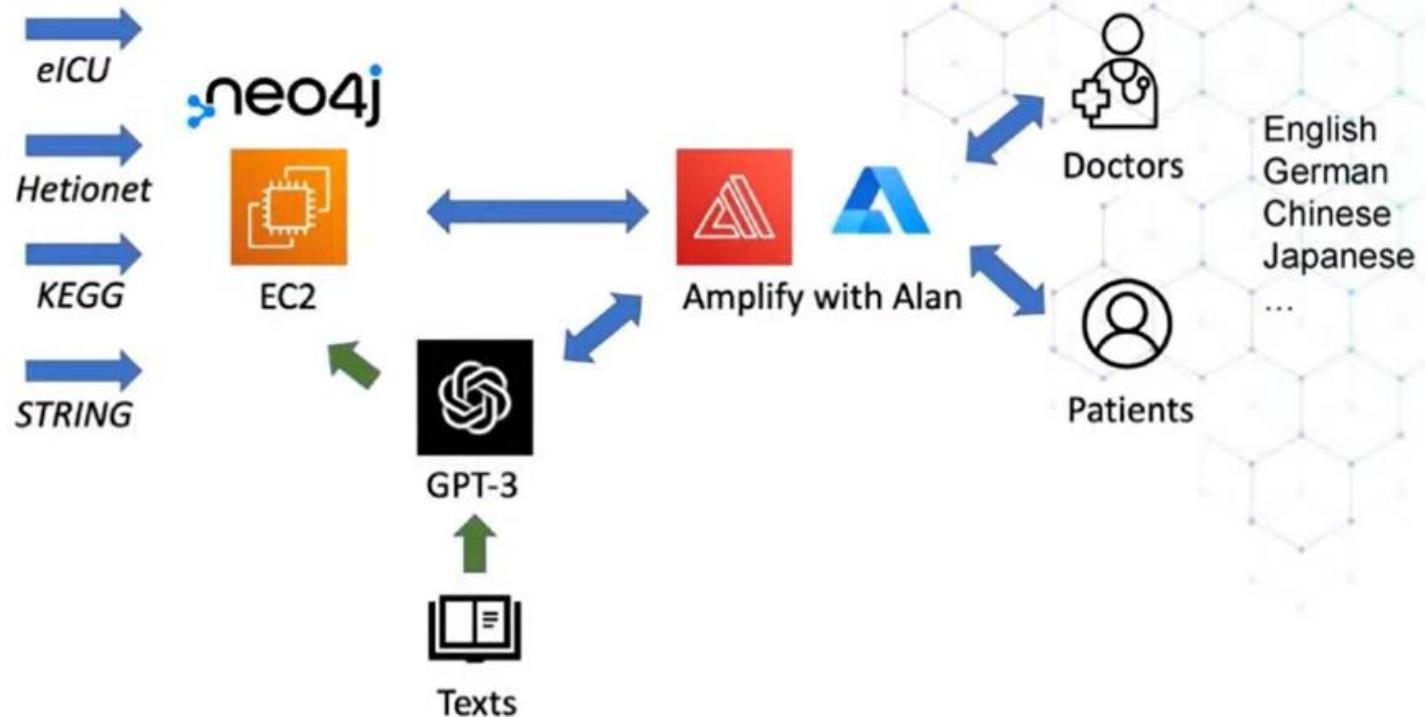
Delicate
Reputation · 2017

Source: Dong, Luna. Building a Broad Knowledge Graph for Products. Keynote at ICDE. 2019

Knowledge Graph in Personalized Medicine Application



The Architecture



GPT-3

Playground

```

    1. Choose this
    code-davinci-001
    Temperature: 0
    Response length: 50
    Stop sequences: Enter sequences and press enter
    Top P: 1
    Frequency penalty: 0
    Presence penalty: 0
    Best of: 1
    Deposit start text: 
    Deposit result text: 
    Show probabilities: Off

    2. Paste your #English-Cypher examples first
    How many times did patient id_3 visit the ED?
    MATCH (p:Patient {patient_id: "id_3"})-->(:VISIT{stay_id})-->(v:PatientVisitStay) RETURN COUNT(v)

    Who did patient id_3 visit the ED?
    MATCH (p:Patient {patient_id: "id_3"})-->(:VISIT{stay_id})-->(v:PatientVisitStay)-->v.visit.visit_by_patient

    Which drug treats COVID-19?
    MATCH (d:Compound)-[:treats]->(c:Disease {name: "COVID-19"}) RETURN d.name

    Which pathogen causes Kyasanur Forest disease?
    MATCH (c:Pathogen)-[:causes]->(d:Disease {name: "Kyasanur Forest disease"}) RETURN c.name

    Which pathogen causes COVID-19?
    MATCH (c:Pathogen)-[:causes]->(d:Disease {name: "COVID-19"}) RETURN c.name

    Which gene causes Christiansson syndrome?
    MATCH (g:Gene)-[:associates]->(d:Disease {name: "Christiansson syndrome"}) RETURN g.name

    Tell me something about the disease named "Christiansson syndrome"
    MATCH (d:Disease {name: "Christiansson syndrome"}) RETURN d.description

    I have Dyspnoea, Hiccup and Edema. What can be the cause of this?
    MATCH (x1:Symptom {name: "Dyspnoea"})-->(:presents)->(d:Disease) MATCH (x2:Symptom {name: "Hiccup"})-->(:presents)->(d:Disease)
    MATCH (x3:Symptom {name: "Edema"})-->(:presents)->(d:Disease) RETURN d.name
  
```

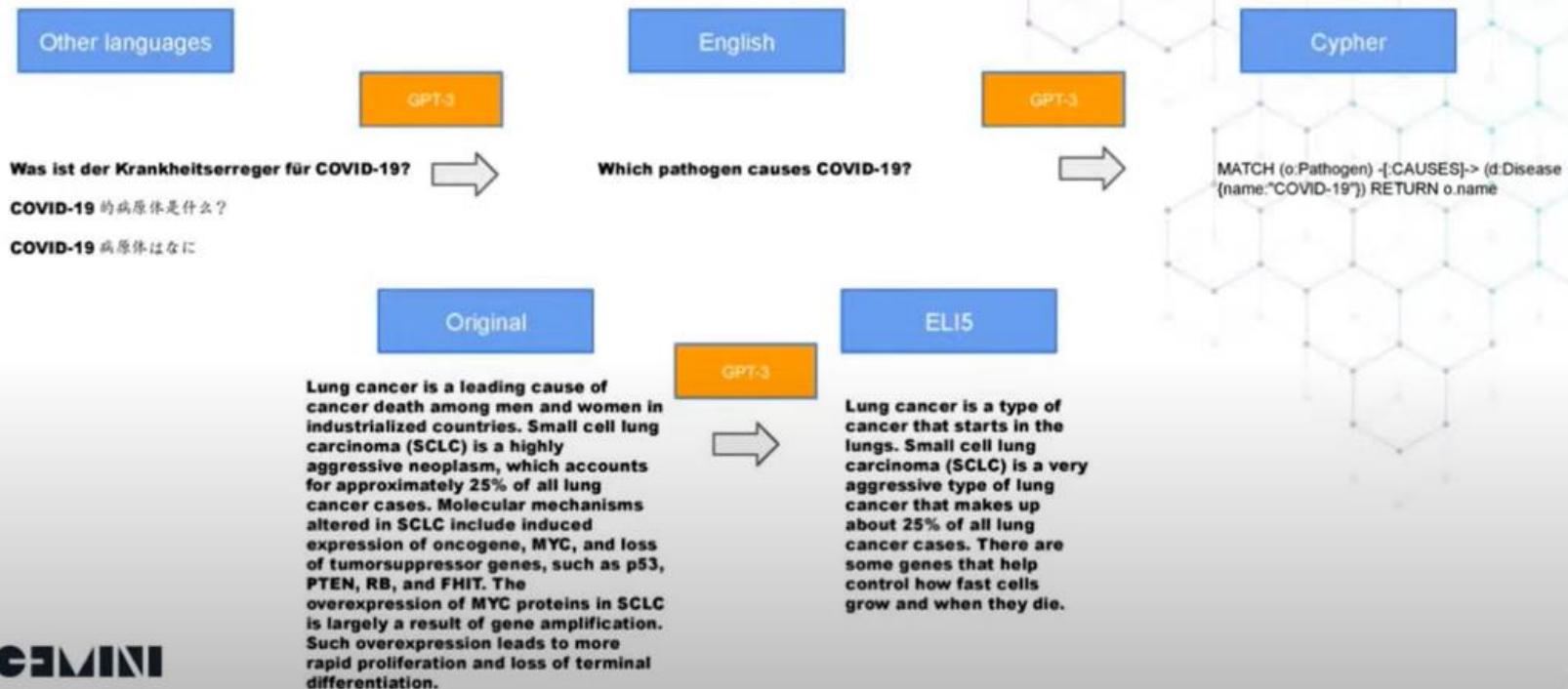
3. Your question

4. Click this

Execute

451 Points

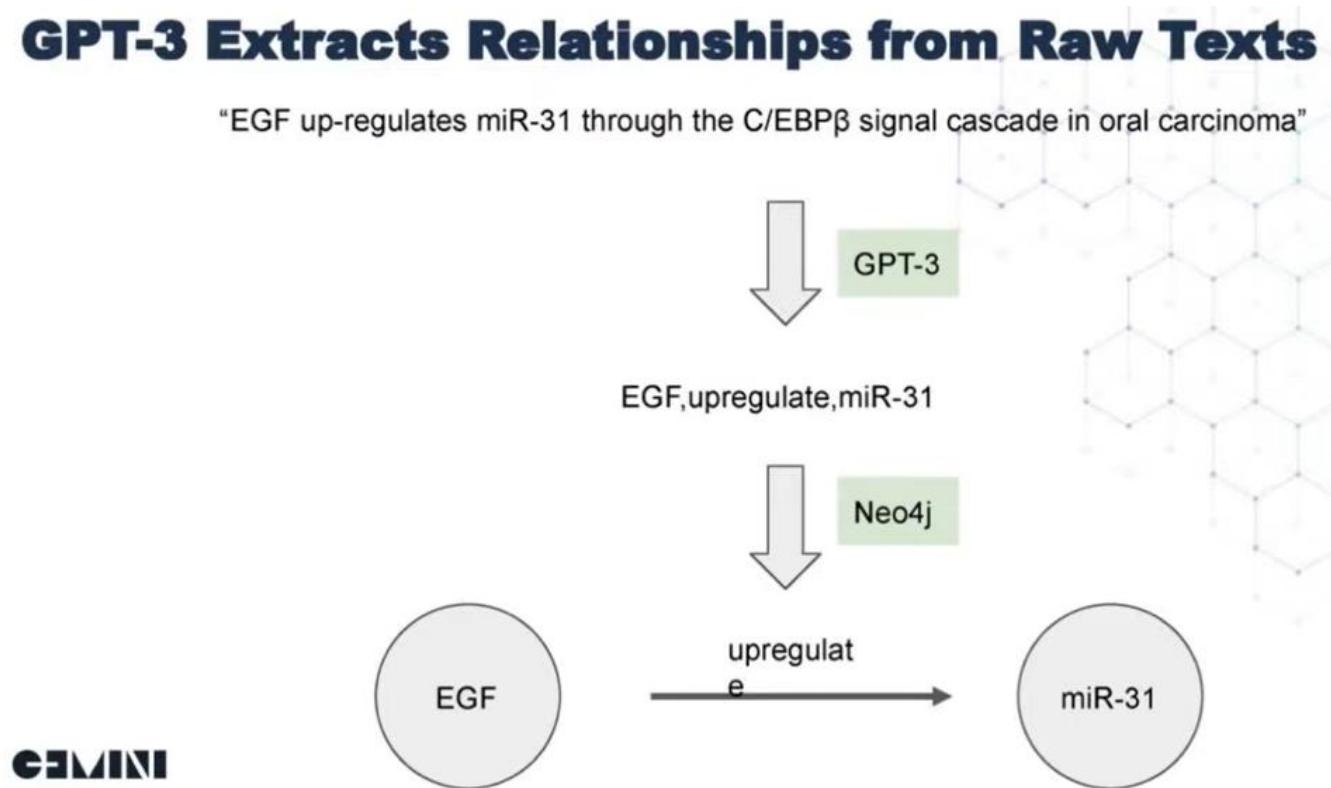
Natural Language Understanding (NLU): GPT-3



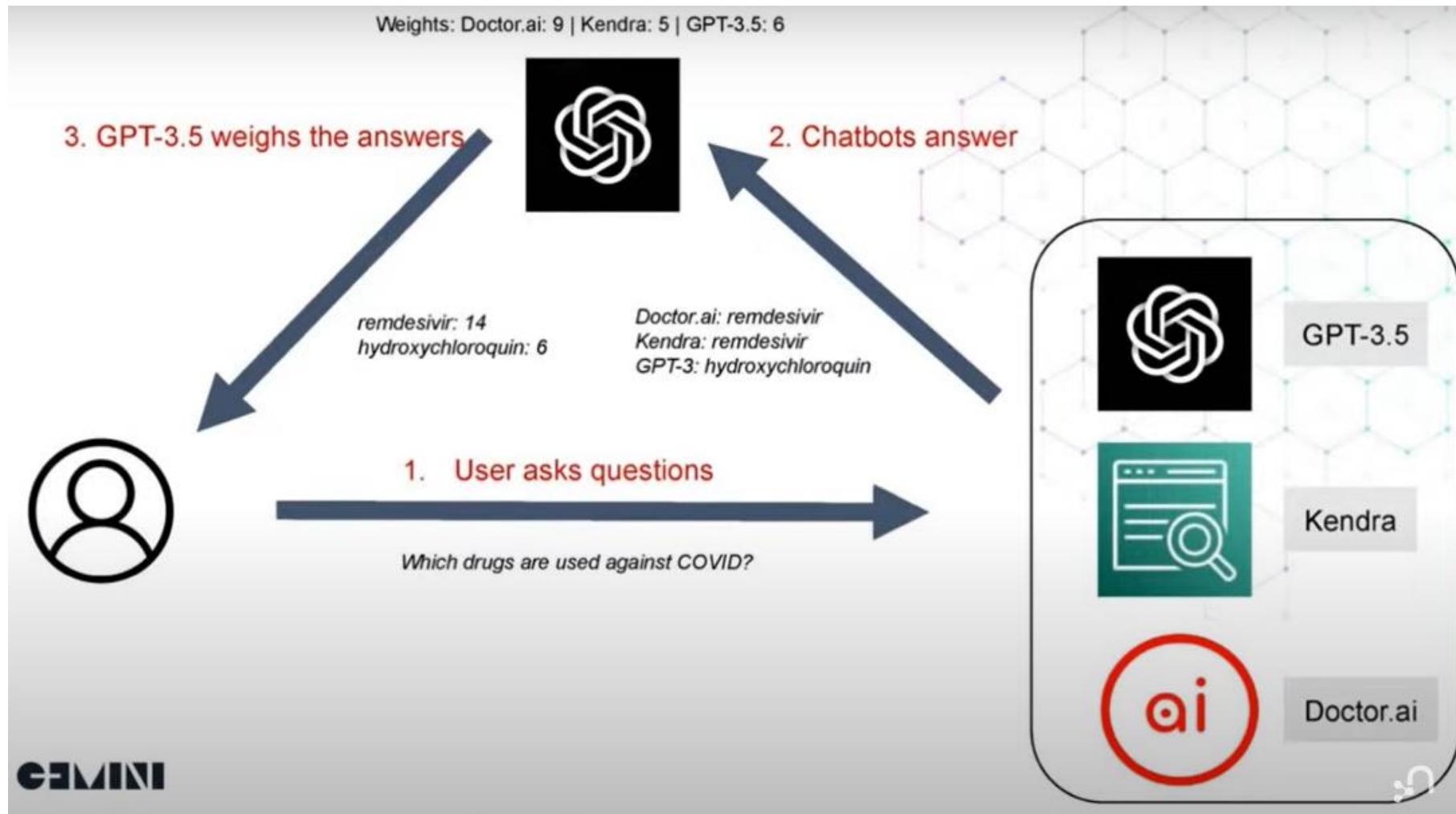
NLP api to extract relations

GPT-3 Extracts Relationships from Raw Texts

"EGF up-regulates miR-31 through the C/EBP β signal cascade in oral carcinoma"

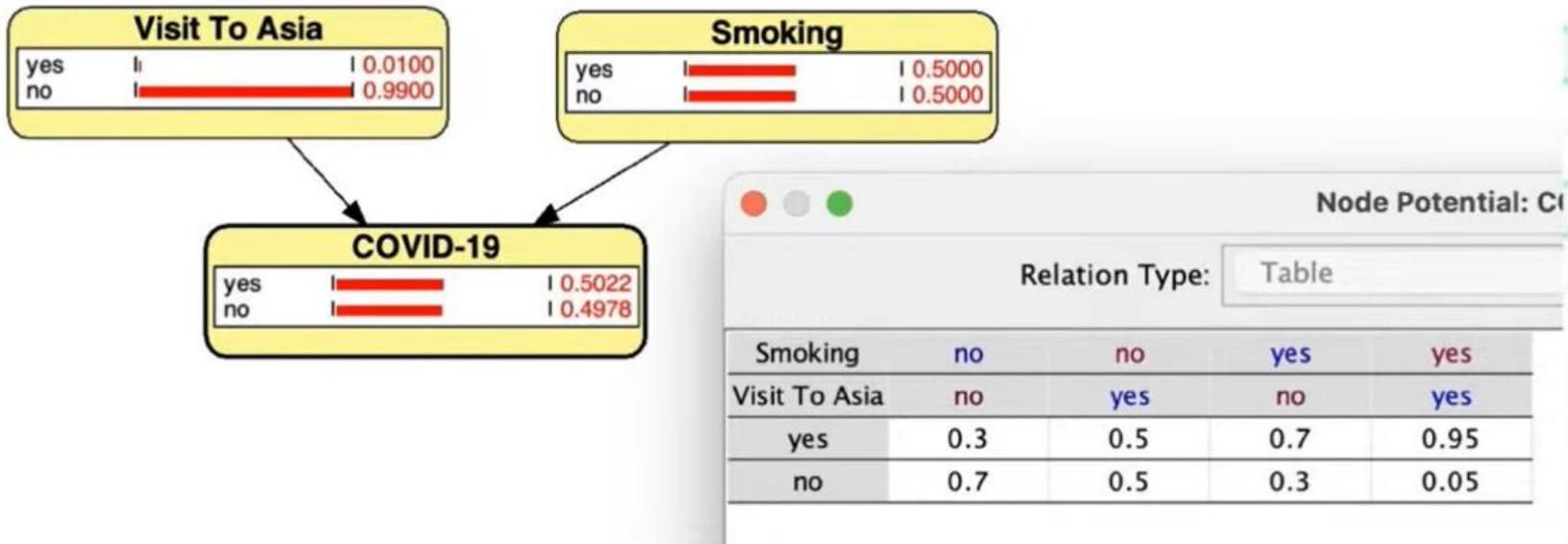


Multiple chatbot output



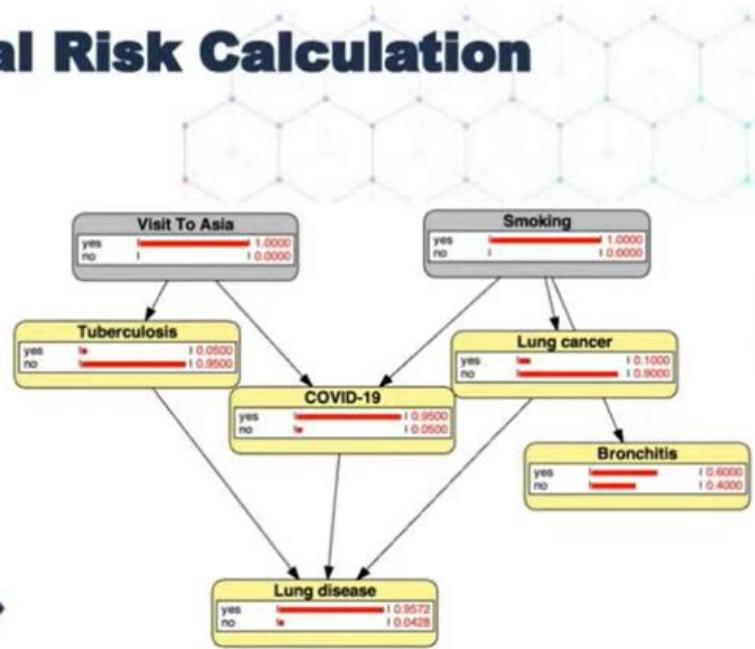
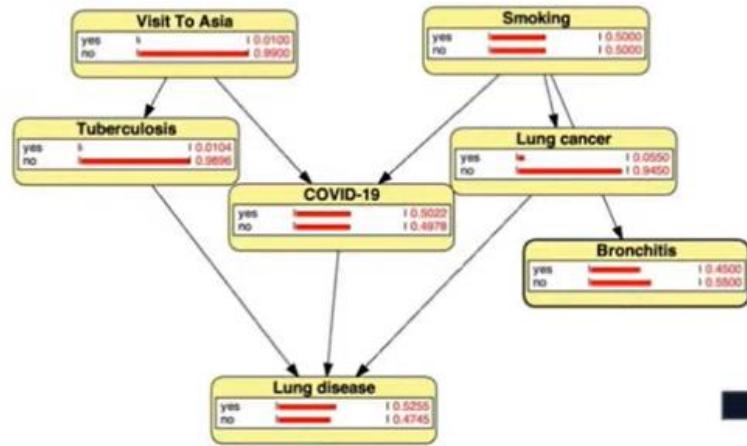
Personalization for medicines

Bayesian Network: A Probabilistic Graph for Decisions

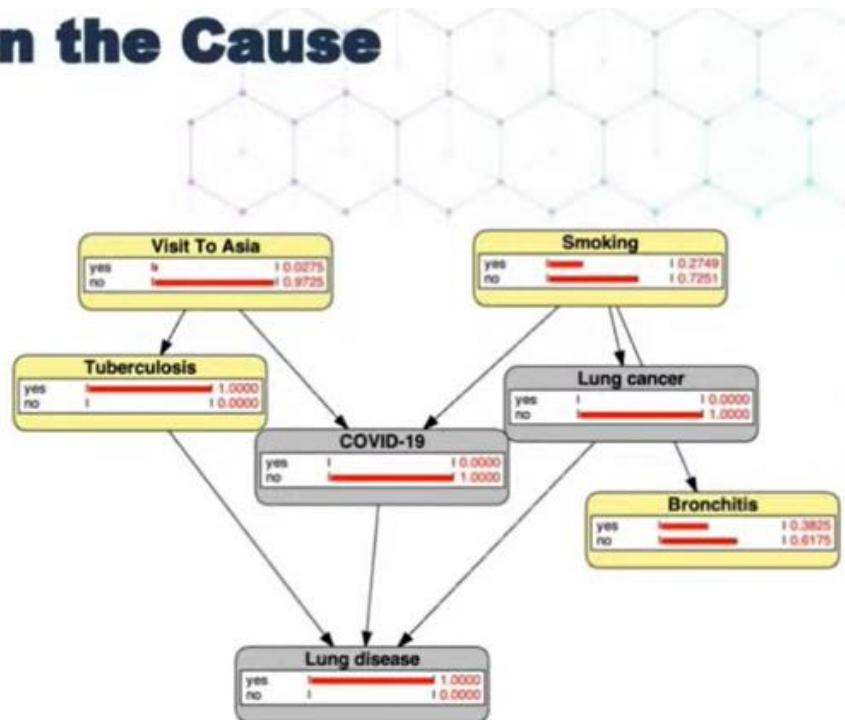
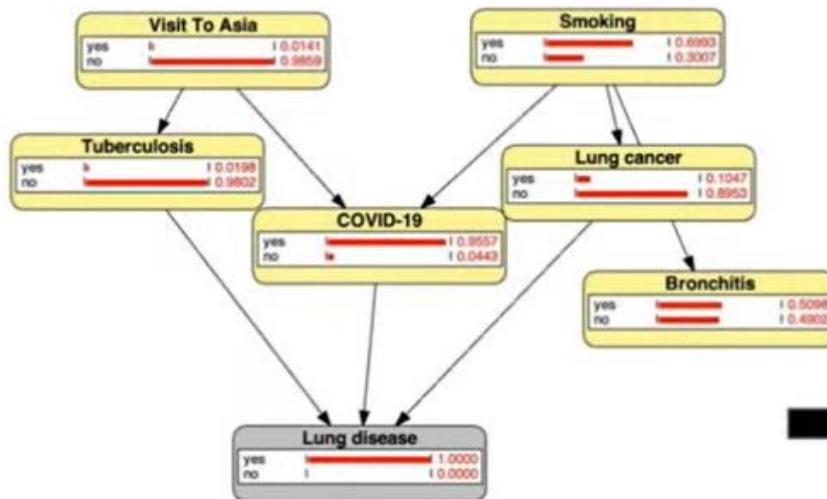


GEMINI

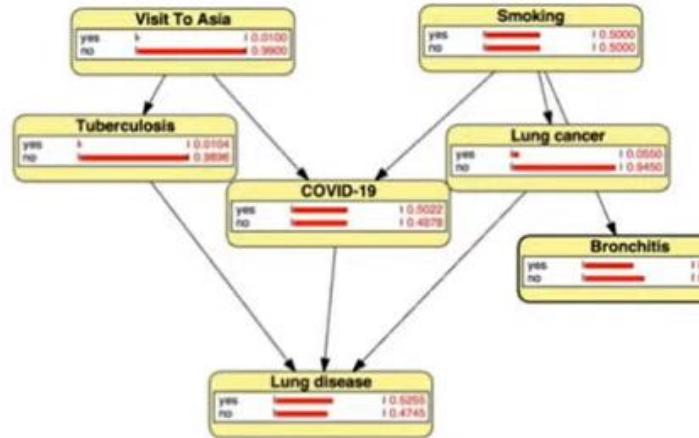
Bayesian Network: Individual Risk Calculation



Bayesian Network: Zero in on the Cause

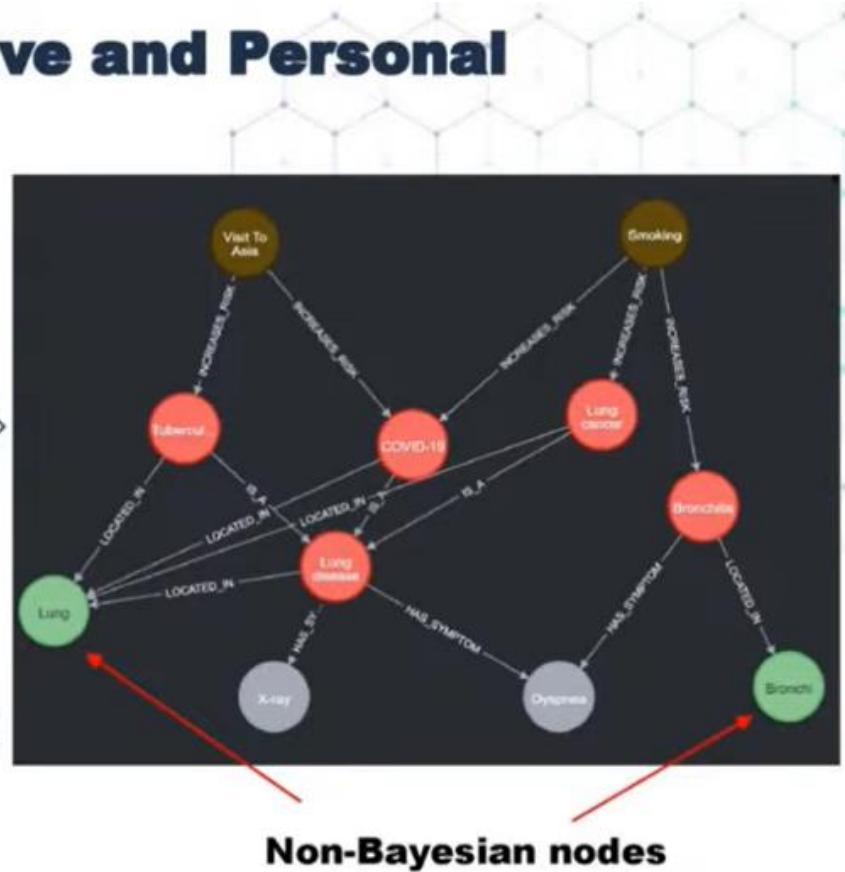


Bayesian KG: Comprehensive and Personal

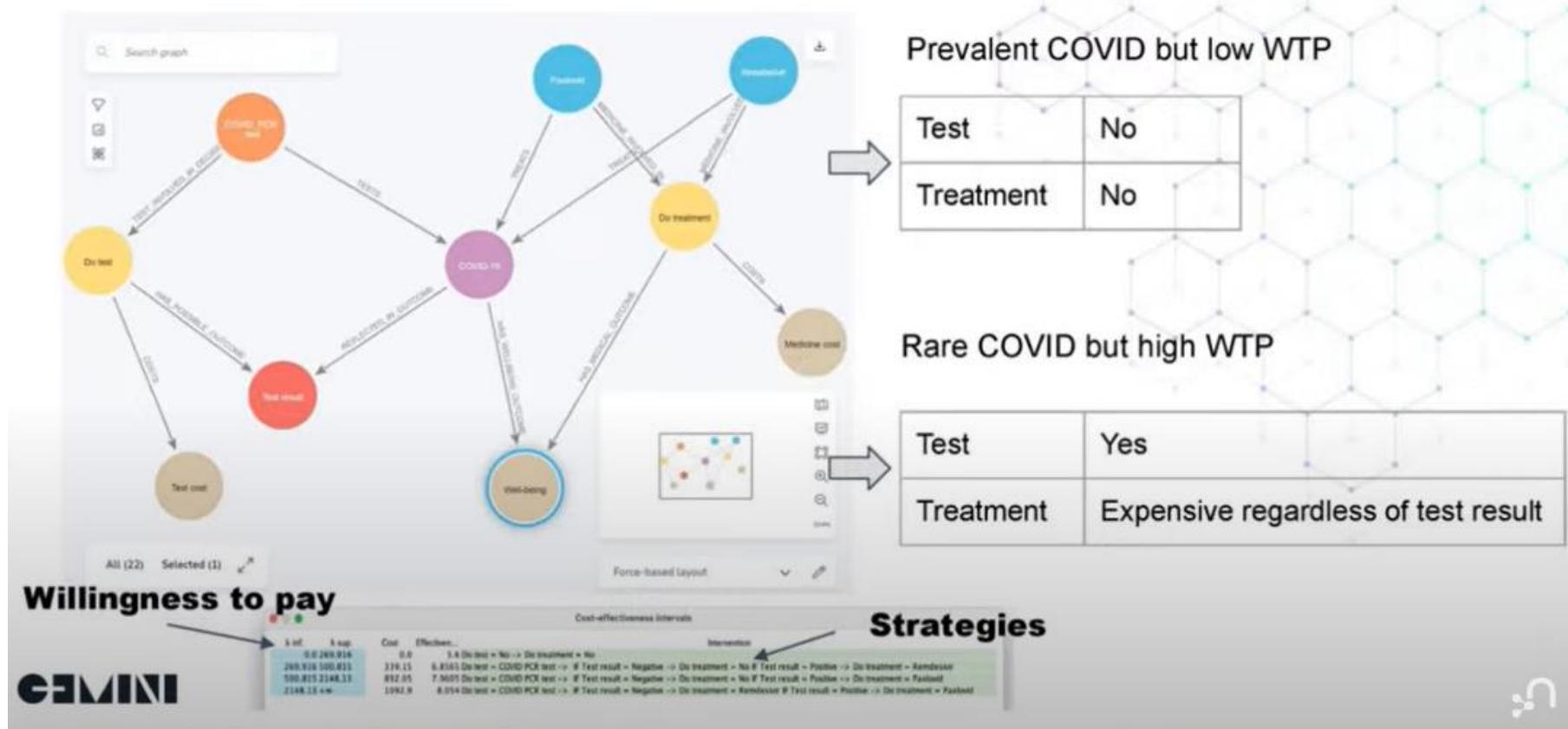


● Non-Bayesian
○ Bayesian

CEMINI

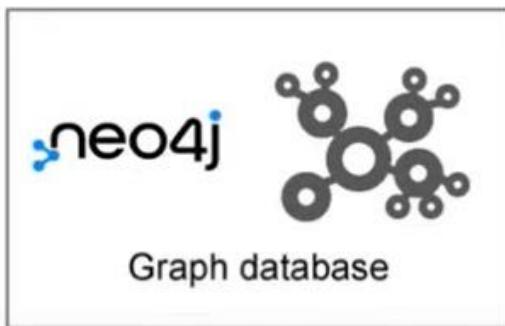


Cost-Benefit: Different Patients Makes Different Decisions

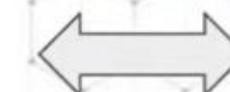
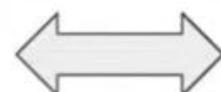




"Who carries out ABC?
Use vector search..."



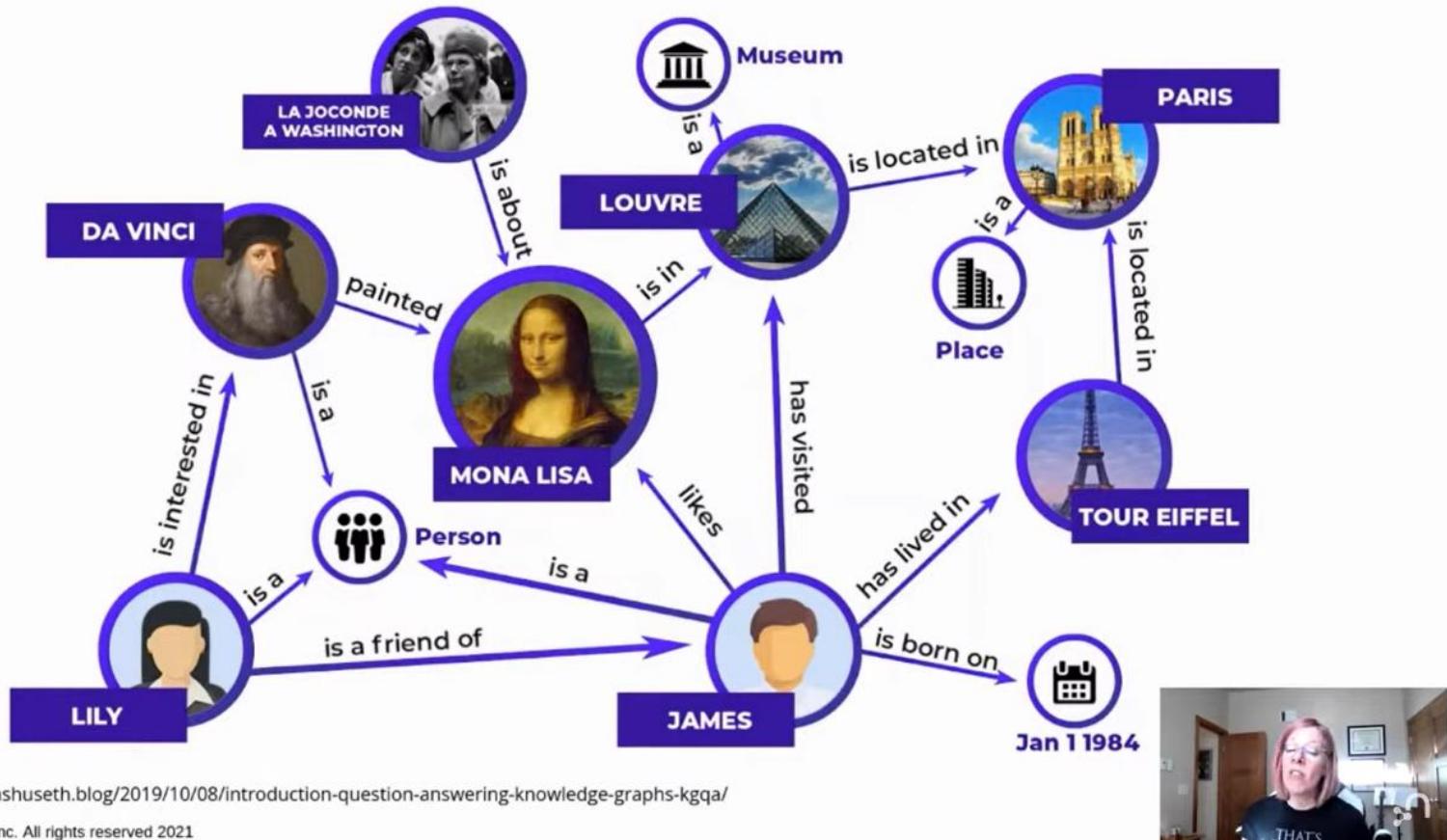
"What does trial
ABC focus on?"



Search the web
about XYZ..."

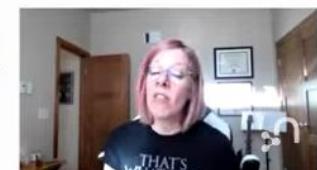


Recommendation system



<https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/>

Neo4j, Inc. All rights reserved 2021



Barack Hussein Obama II PERSON ((listen) bo-RAHK hoo-SAYN oh-BAH-ma; born August 4, 1961 DATE) is an American NORP politician and attorney who served as the 44th ORDINAL president of the United States GPE from 2009 DATE to 2017 DATE. A member of the Democratic Party ORG, Obama PERSON was the first ORDINAL African-American NORP president of the United States GPE. He previously served as a U.S. GPE senator from Illinois GPE from 2005 to 2008 and as an Illinois GPE state senator from 1997 DATE to 2004.

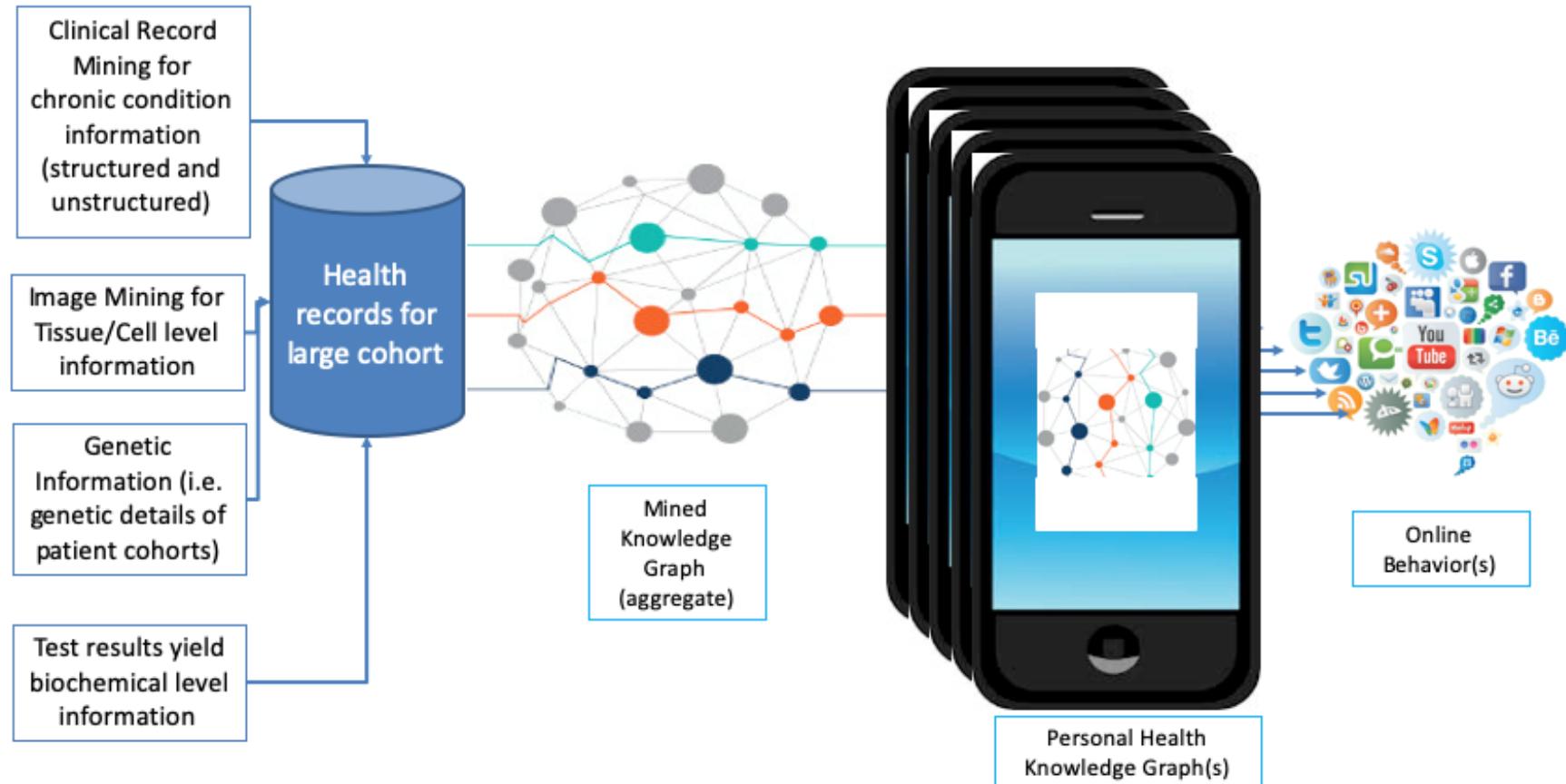
Obama PERSON was born in Honolulu GPE, Hawaii GPE. After graduating from Columbia University ORG in 1983 DATE, he worked as a community organizer in Chicago GPE. In 1988 DATE, he enrolled in Harvard Law School ORG, where he was the first ORDINAL black person to be president of the Harvard Law Review ORG. After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School ORG from 1992 DATE to 2004. Turning to elective politics, he represented the 13th ORDINAL district from 1997 DATE until 2004 DATE in the Illinois Senate, when he ran for the U.S. Senate ORG. Obama PERSON received national attention in 2004 DATE with his March Senate ORG primary win, his well-received July DATE Democratic National Convention keynote address, and his landslide November DATE election to the Senate ORG. In 2008 DATE, he was nominated by the Democratic Party ORG for president a year DATE after beginning his campaign, and after a close primary campaign against Hillary Clinton PERSON. Obama PERSON was elected over Republican NORP Senator John McCain PERSON in the general election and was inaugurated alongside his running mate, Joe Biden PERSON, on January 20, 2009 DATE. Nine months later DATE, he was named the 2009 DATE Nobel Peace Prize WORK_OF_ART laureate.

Obama PERSON signed many landmark bills into law during his first two years DATE in office. The main reforms that were passed include the Affordable Care Act LAW (commonly referred to as ACA ORG or "Obamacare WORK_OF_ART"), although without a public health insurance option, the Dodd-Frank Wall Street Reform and Consumer Protection Act, and the Don't Ask, Don't Tell Repeal Act of 2010 DATE. The American Recovery and Reinvestment Act ORG of 2009 DATE and Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010 DATE served as economic stimuli amidst the Great Recession EVENT. After a lengthy debate over the national debt limit, he signed the Budget Control ORG and the American Taxpayer Relief Act ORG. In foreign policy, he increased U.S. GPE troop levels in Afghanistan GPE, reduced nuclear weapons with the United States- GPE Russia New START treaty, and ended military involvement in the Iraq War EVENT. He ordered military involvement in Libya GPE for the implementation of the UN Security Council ORG Resolution 1973 DATE, contributing to the overthrow of Muammar Gaddafi PERSON. He also ordered the military operations that resulted in the deaths of Osama bin Laden PERSON and suspected American NORP Al-Qaeda ORG operative Anwar al-Awlaki PERSON.

After winning re-election by defeating Republican NORP opponent Mitt Romney PERSON, Obama PERSON was sworn in for a second ORDINAL term in 2013 DATE. During this term, he promoted inclusion for LGBT Americans NORP. His administration filed briefs that urged the Supreme Court ORG to strike down same-sex marriage bans as unconstitutional (United States GPE v. Windsor PERSON and Obergefell ORG v. Hodges PERSON); same-sex marriage was legalized nationwide in 2015 DATE after the Court ORG ruled so in Obergefell ORG. He advocated for gun control in response to the Sandy Hook Elementary School ORG shooting, indicating support for a ban on assault weapons, and issued wide-ranging executive actions concerning global warming and immigration. In foreign policy, he ordered military intervention in Iraq GPE in response to gains made by ISIL ORG after the 2011 DATE withdrawal from Iraq GPE, continued the process of ending U.S. GPE combat operations in Afghanistan GPE in 2016 DATE, promoted discussions that led to the 2015 DATE Paris Agreement EVENT on global climate change, initiated sanctions against Russia GPE following the invasion in Ukraine GPE and again after interference in the 2016 DATE U.S. GPE elections, brokered the JCPOA ORG nuclear deal with Iran GPE, and normalized U.S. GPE relations with Cuba GPE. Obama PERSON nominated three CARDINAL justices to the Supreme Court ORG: Sonia Sotomayor PERSON and Elena Kagan PERSON were confirmed as justices, while Merrick Garland PERSON faced partisan obstruction from the Republican NORP-led Senate ORG led by Mitch McConnell PERSON, which never held hearings or a vote on the nomination. Obama PERSON left office in January 2017 DATE after Washington GPE, D.C. During Obama's PERSON term in office, the United States GPE reputation abroad, as well as the American NORP economy, significantly improved. Obama PERSON's presidency has generally been viewed favorably, and evaluations of his presidency among historians, political scientists, and the general public frequently place him among the upper tier of American NORP presidents.



Health Knowledge Graph



Knowledge Graphs for Social Good (KGSG)

Best Practices, Methods, and Challenges - Held May 4th, 2020 at KGC 2020

≡ accenture

Technology Innovation

Applying knowledge graphs for social good

JUNE 26, 2020

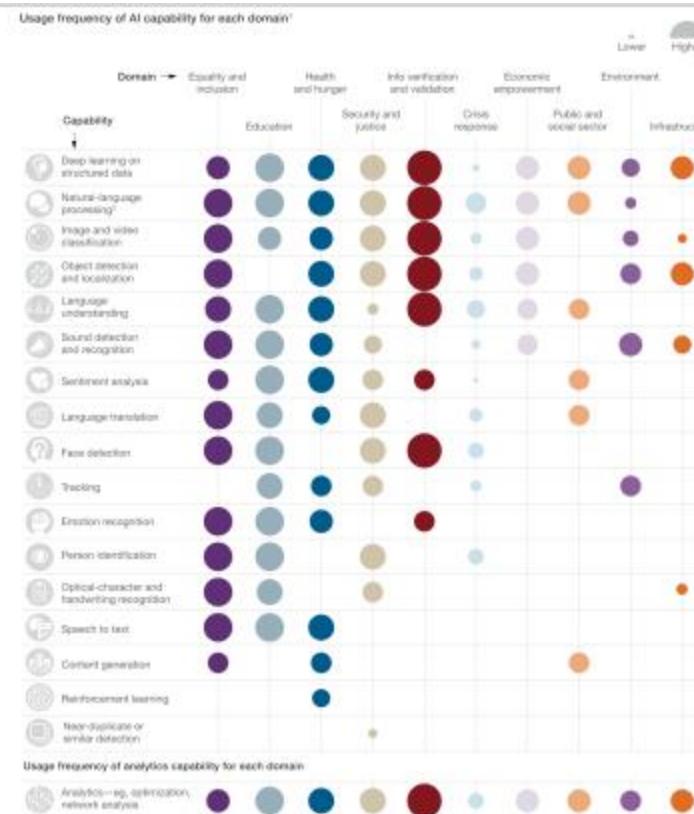
Knowledge Graphs for Social Good: An Entity-centric Search Engine for the Human Trafficking Domain

Publisher: IEEE

Cite This

PDF

Mayank Kejriwal · Pedro Szekely · All Authors



KG in explainable AI

Explainable AI

On The Role of Knowledge Graphs in Explainable AI

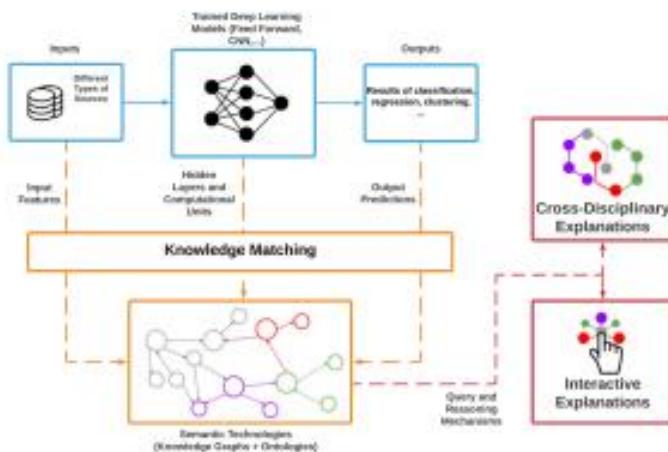
Freddy Lecue^{a,b}

^a CortAlx, Thales, Montreal, Canada

E-mail: freddy.lecure@inria.fr

^b WIMMICS, Inria, Sophia Antipolis, France

E-mail: freddy.lecure@thalesgroup.fr



Explainable Artificial Intelligence (XAI)

Dr. Matt Turek

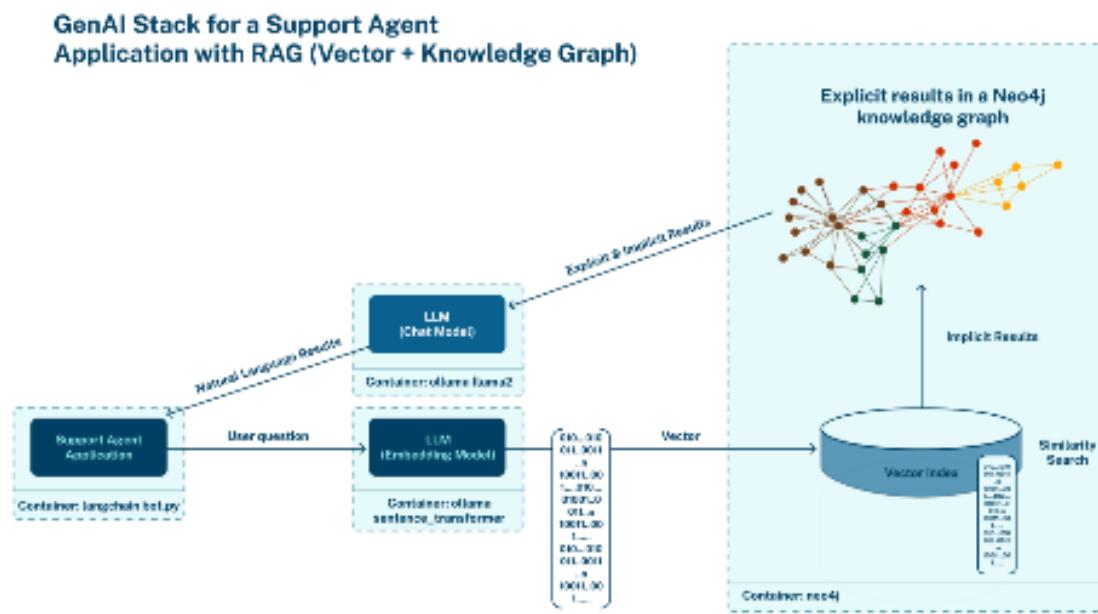


Source: Knowledge Graphs For eXplainable AI. On the Integration of Semantic Technologies and Symbolic Systems into Deep Learning Models for a More Comprehensible Artificial Intelligence.

<https://towardsdatascience.com/knowledge-graphs-for-explainable-ai-dcd73c5c016>

KG for grounding LLM

- KG for grounding LLMs due to ability to represent both structured and unstructured data, [unlike vector databases](#).
 - Technique is called [retrieval augmented generation \(RAG\)](#)
 - LLM *retrieves* relevant information from the knowledge graph using vector and semantic search. Then *augments* the response with the contextual data in the knowledge graph.
 - This RAG LLM process generates more precise, accurate, and contextually relevant output, while preventing false information, also known as *LLM hallucination*.



Scientific Text Mining

Jiang, M., & Shang, J. (2020, August). Scientific Text Mining and Knowledge Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3537-3538).

Question Answering

Hixon, B., Clark, P., & Hajishirzi, H. (2015). Learning knowledge graphs for question answering through conversational dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 851-861).

Recommendation Systems

Oramas, S., Ostuni, V. C., Noia, T. D., Serra, X., & Sciascio, E. D. (2016). Sound and music recommendation with knowledge graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(2), 1-21.

Summarization

Gunaratna, K., Yazdavar, A. H., Thirunarayan, K., Sheth, A., & Cheng, G. (2017, August). Relatedness-based multi-entity summarization. In *IJCAI: proceedings of the conference* (Vol. 2017, p. 1060). NIH Public Access.

Truth/fact-checking

Shiralkar, P., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2017, November). Finding streams in knowledge graphs to support fact checking. In *2017 IEEE International Conference on Data Mining (ICDM)* (pp. 859-864). IEEE.

References

<https://slideplayer.com/slide/13562607/>

[displaCy Named Entity Visualizer · Explosion](#)

Knowledge Graph for Personalized Medicines Application

<https://www.youtube.com/watch?v=nbY0KoT54uE>

<https://www.youtube.com/watch?v=BOYurPHWWXs>

References

- <https://www.slideshare.net/STI-Innsbruck/knowledge-graphs-236706310>
 - <https://www.slideshare.net/TakanoriUgai/practical-use-of-knowledge-graph-with-case-studies-using-semantics-web-publishing-tools>
 - <https://www.slideshare.net/elenasimperl/knowledge-graph-use-cases-in-natural-language-generation>
 - <https://redirect.cs.umbc.edu/courses/graduate/691/fall18/07/resources.html>
-



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 9: Statistical Machine translation

Date – 11th February 2024

Time – 1.40 pm to 3.40 pm

These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philipp Koehn**, Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

Agenda

Statistical Machine translation

- Introduction
- Approaches
- Parallel Corpora
- Word Alignment
- Language Models
- Translation Models
- IBM Models
- MT Evaluation
- Bleu Score

Machine Translation



Machine Translation (MT) is the task of translating a sentence x from one language (the *source language*) to a sentence y in another language (the *target language*).

x: *L'homme est né libre, et partout il est dans les fers*



y: *Man is born free, but everywhere he is in chains*

- Rousseau

Machine Translation

- Automatically translate one natural language into another.

Mary didn't slap the green witch.



Maria no dió una bofetada a la bruja verde.

Ambiguity Resolution is Required for Translation



- Syntactic and semantic ambiguities must be properly resolved for correct translation:
 - “John plays the guitar.” → “John toca la guitarra.”
 - “John plays soccer.” → “John juega el fútbol.”
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
 - “The spirit is willing but the flesh is weak.” ⇒ “The liquor is good but the meat is spoiled.”
 - “Out of sight, out of mind.” ⇒ “Invisible idiot.”

Ambiguity Resolution is Required for Translation

He deposited money in a bank account
with a high interest rate.

Sitting on the bank of the Mississippi,
a passing ship piqued his interest.

1950s: Early Machine Translation



*Machine Translation research began in the **early 1950s**.*

- *Russian → English
(motivated by the Cold War!)*
- *Systems were mostly **rule-based**, using a bilingual dictionary to map Russian words to their English counterparts*



1 minute video showing 1954 MT:
<https://youtu.be/K-HfpsHPmvw>

Rule Based Systems

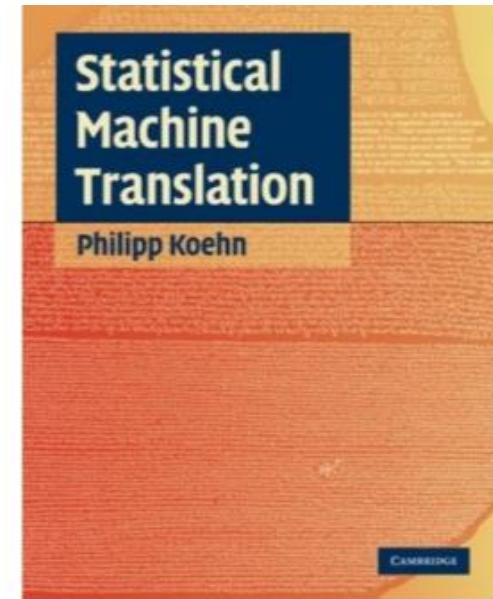
- Rule-based systems
 - build dictionaries
 - write transformation rules
 - refine, refine, refine
- Météo system for weather forecasts (1976)
- Systran (1968), Logos and Metal (1980s)

```
"have" :=  
  
if  
    subject/animate)  
    and object/owned-by-subject)  
then  
    translate to "kade... aahe"  
if  
    subject/animate)  
    and object/kinship-with-subject)  
then  
    translate to "laa... aahe"  
if  
    subject/inanimate)  
then  
    translate to "madhye...  
aahe"
```

Statistical Machine Translation



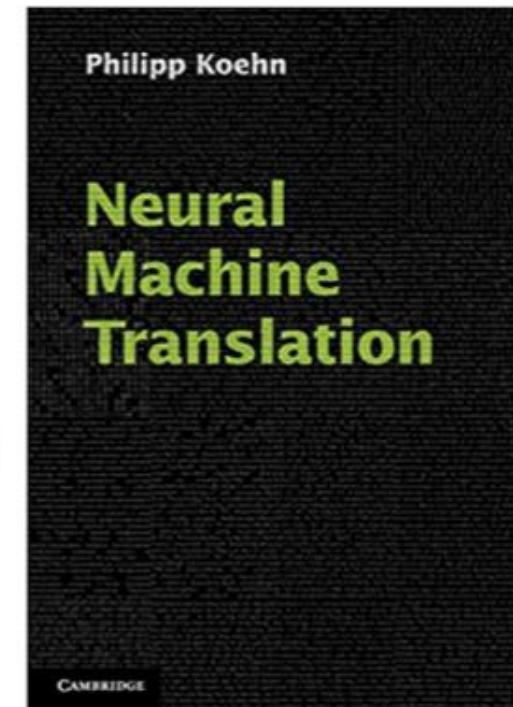
- 1980s: IBM
- 1990s: increased research
- Mid 2000s: Phrase-Based MT (Moses, Google)
- Around 2010: commercial viability



Neural Machine Translation



- Late 2000s: neural models for computer vision
- Since mid 2010s: neural models for machine translation
- 2016: Neural machine translation the new state of the art



Linguistic Issues Making MT Difficult



- Languages have different sentence structure
- Syntactic variation between **SVO** (e.g. English), **SOV** (e.g. Hindi), and **VSO** (e.g. Arabic) languages.
 - SVO languages use prepositions
 - SOV languages use postpositions
- Morphological issues for languages with complex word structure.
- **Pro-drop** ("pronoun-dropping") languages regularly omit subjects that must be inferred.

तुमने नाद्या को खाना दिया?
Did you give the food to Nadya

हाँ दे दिया।
yes, (*I*) gave (*her* food).

Semantic Ambiguity

Pronominal anaphora

I saw the movie and it is good.

- How to translate it into German (or French)?
 - it refers to movie
 - movie translates to Film
 - Film has masculine gender
 - ergo: it must be translated into masculine pronoun er

Semantic Ambiguity

Discourse

Since you brought it up, I do not agree with you.

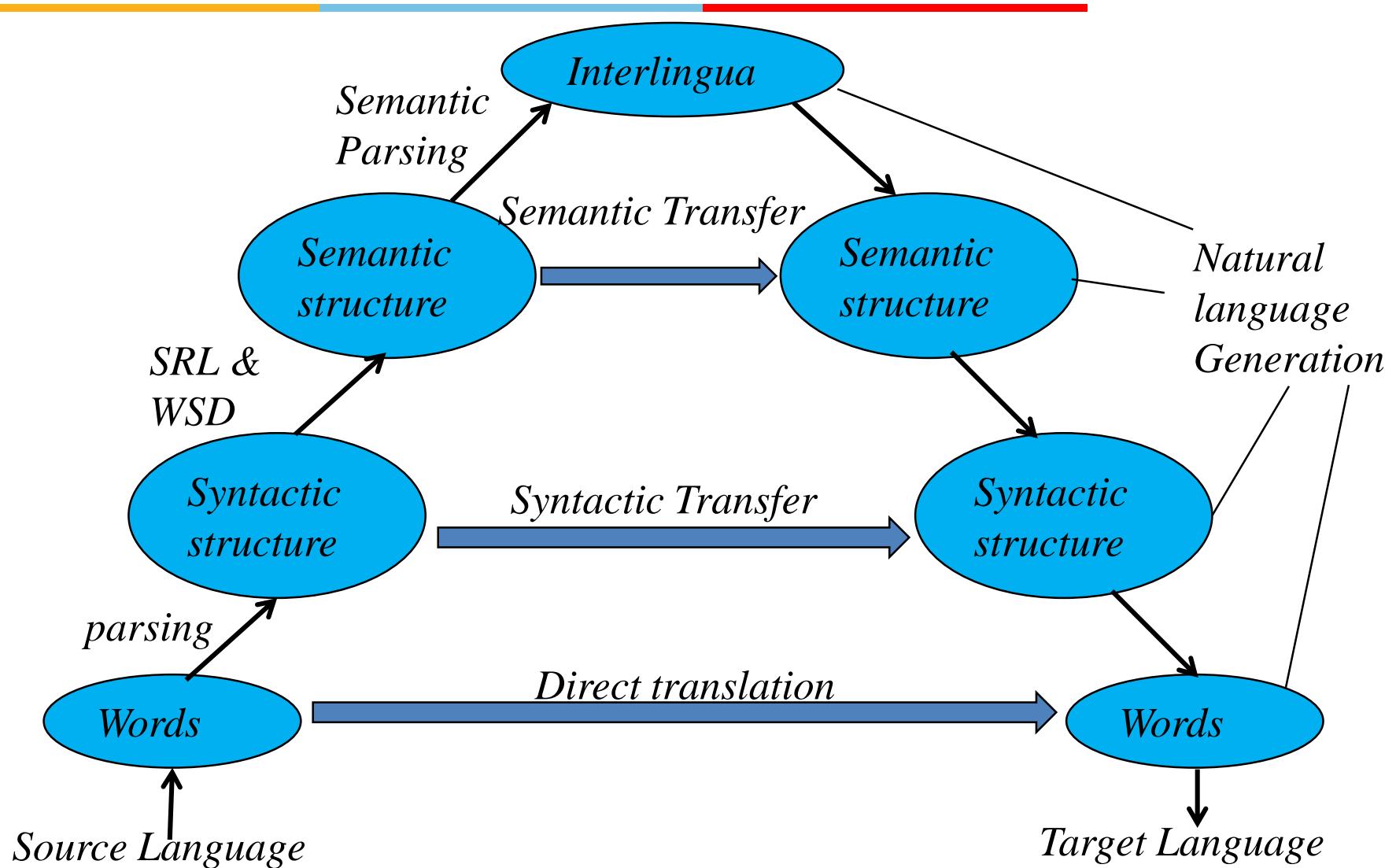
Since you brought it up, we have been working on it.

- How to translated since? Temporal or conditional?
 - Analysis of discourse structure — a hard problem
-

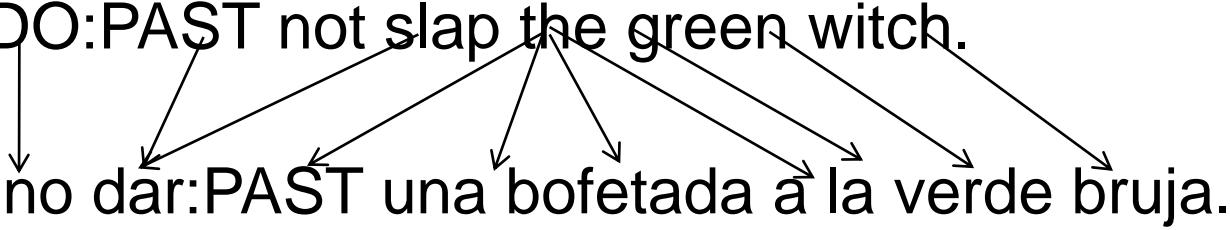
Lexical Gaps

- Some words in one language do not have a corresponding term in the other.
 - Rivière (river that flows into ocean) and fleuve (river that does not flow into ocean) in French
 - Schedenfraude (feeling good about another's pain) in German.
 - Oyakoko (filial piety- virtue of respect for one's parents, elders) in Japanese

Vauquois Triangle



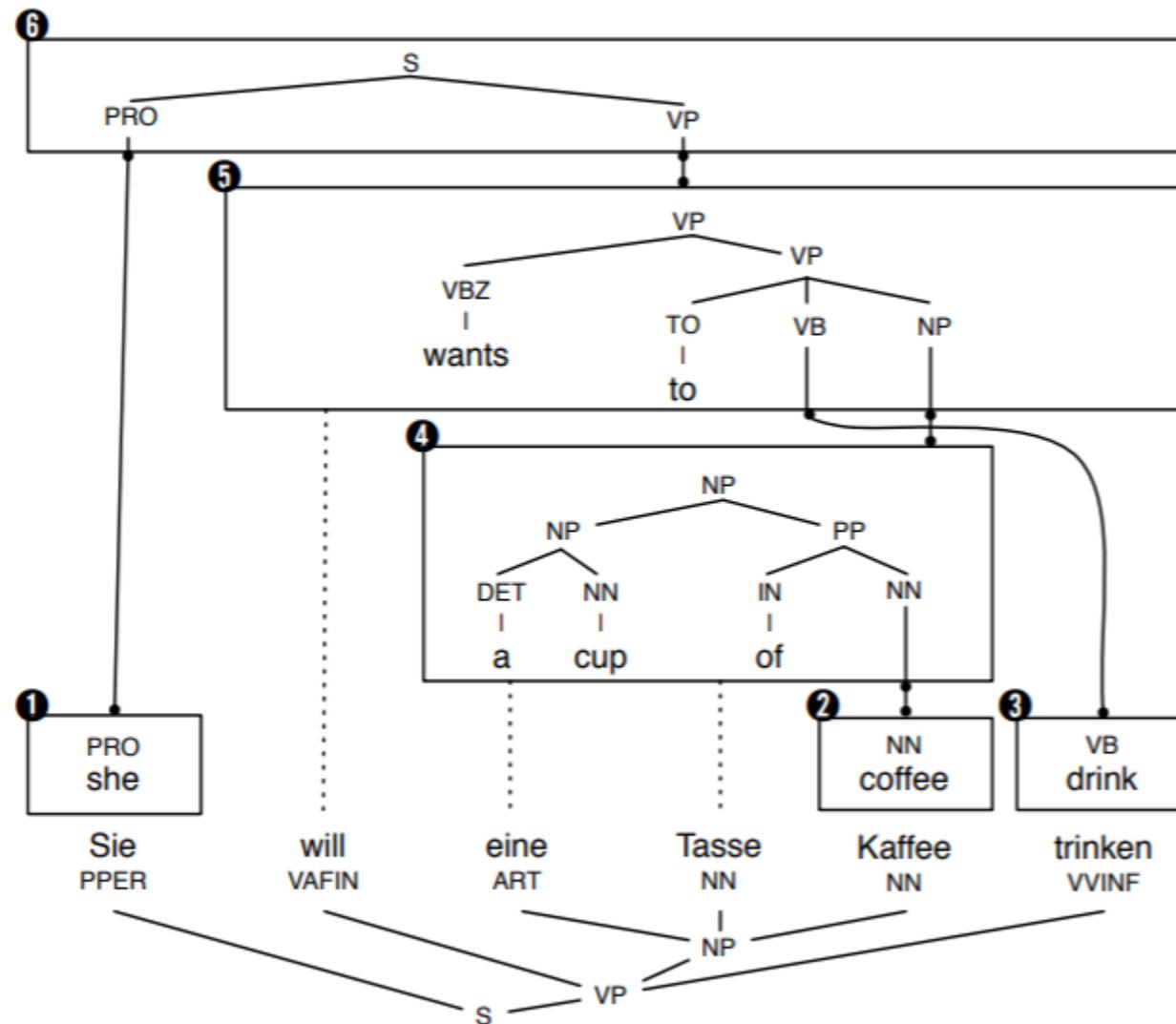
Direct Transfer

- Morphological Analysis
 - Mary didn't slap the green witch. →
Mary DO:PAST not slap the green witch.
- Lexical Transfer
 - Mary DO:PAST not slap the green witch.
 - Maria no dár:PAST una bofetada a la verde bruja.
- Lexical Reordering
 - Maria no dar:PAST una bofetada a la bruja verde.
- Morphological generation
 - Maria no dió una bofetada a la bruja verde.

Syntactic Transfer

- Simple lexical reordering does not adequately handle more dramatic reordering such as that required to translate from an SVO to an SOV language.
- Need syntactic transfer rules that map parse tree for one language into one for another.
 - English to Spanish:
 - $NP \rightarrow Adj\ Nom \Rightarrow NP \rightarrow Nom\ ADJ$
 - English to Japanese:
 - $VP \rightarrow V\ NP \Rightarrow VP \rightarrow NP\ V$
 - $PP \rightarrow P\ NP \Rightarrow PP \rightarrow NP\ P$

Syntax-based Translation



Semantic Transfer

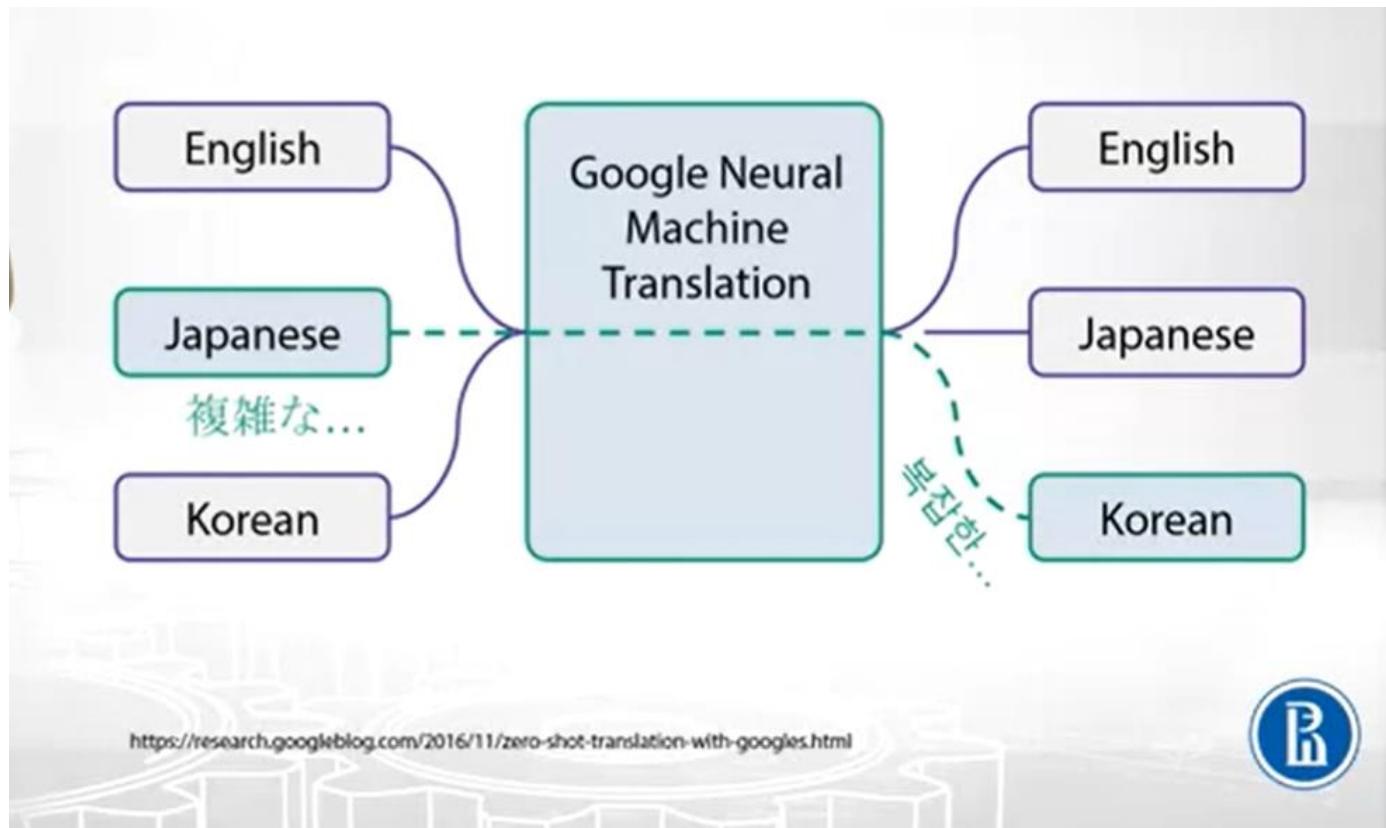
- Some transfer requires semantic information.
- Semantic roles can determine how to properly express information in another language.
- In Chinese, PPs that express a goal, destination, or benefactor occur *before* the verb but those expressing a recipient occur *after* the verb.
- Transfer Rule
 - English to Chinese
 - $VP \rightarrow V\ PP[+\text{benefactor}] \Rightarrow VP \rightarrow PP[+\text{benefactor}] V$

Semantic Transfer

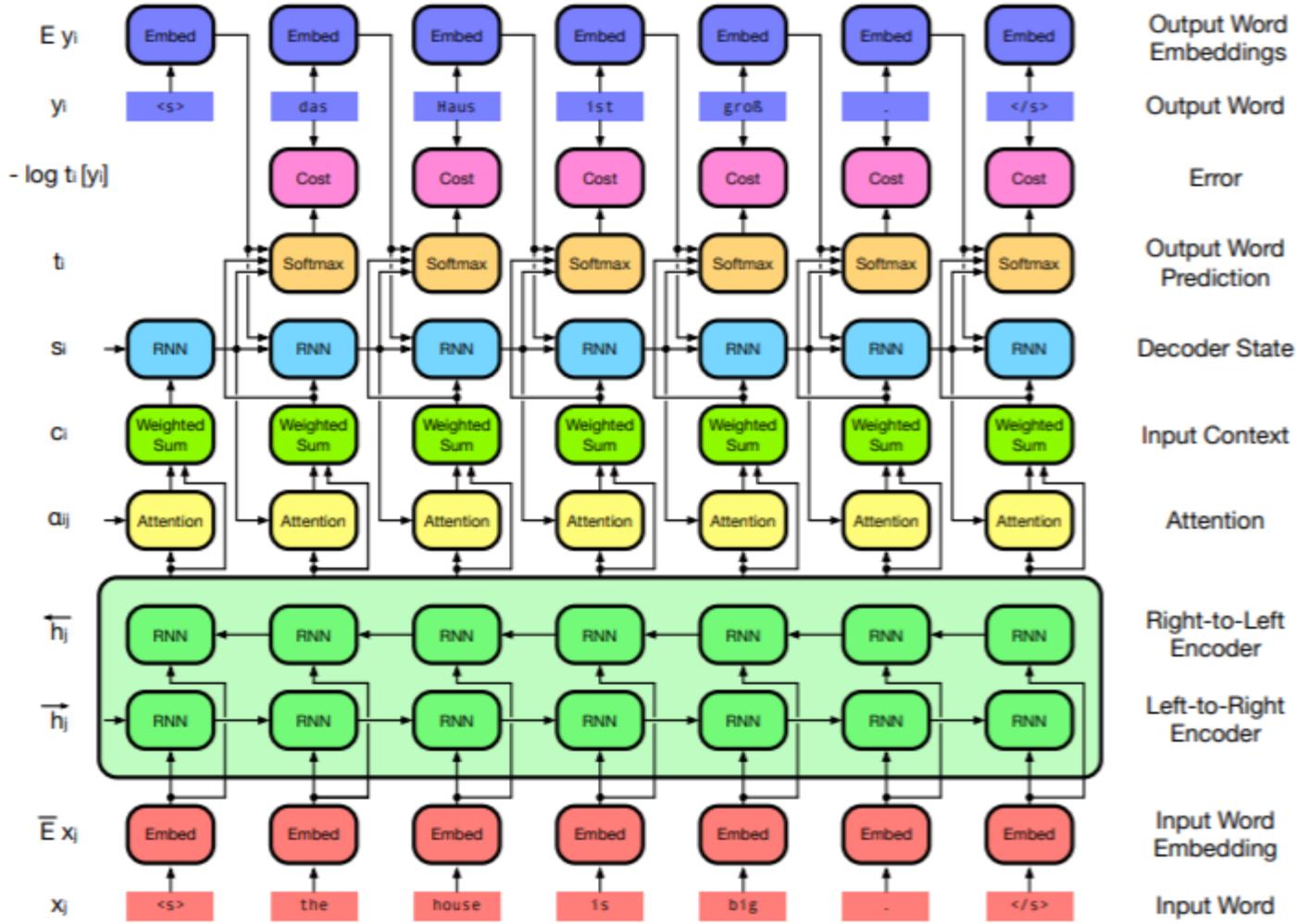
- Generalizes over equivalent syntactic constructs
(e.g., active and passive)

- Defines semantic relationships
 - semantic roles
 - co-reference
 - discourse relations

Interlingua using neural network



Neural Network



Statistical MT

- Manually encoding comprehensive bilingual lexicons and transfer rules is difficult.
- SMT acquires knowledge needed for translation from a ***parallel corpus*** or ***bitext*** that contains the same set of documents in two languages.
- The Canadian Hansards (parliamentary proceedings in French and English) is a well-known parallel corpus.
- First align the sentences in the corpus based on simple methods that use coarse cues like sentence length to give bilingual sentence pairs.

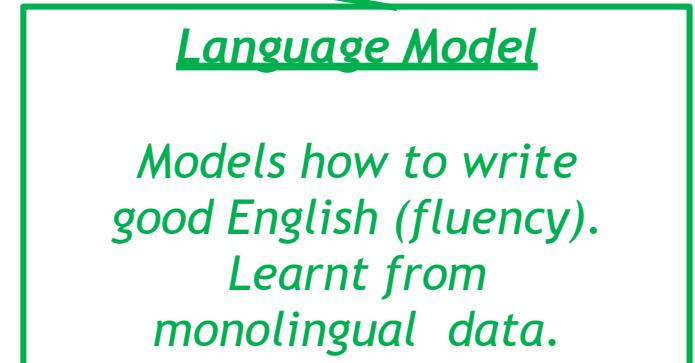
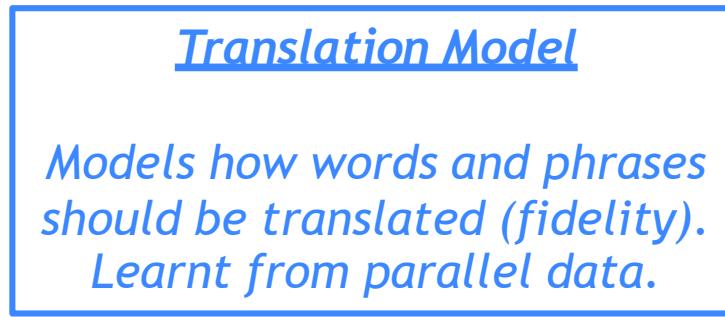
<https://github.com/joshua-decoder/indian-parallel-corpora>

1990s-2010s: Statistical Machine Translation

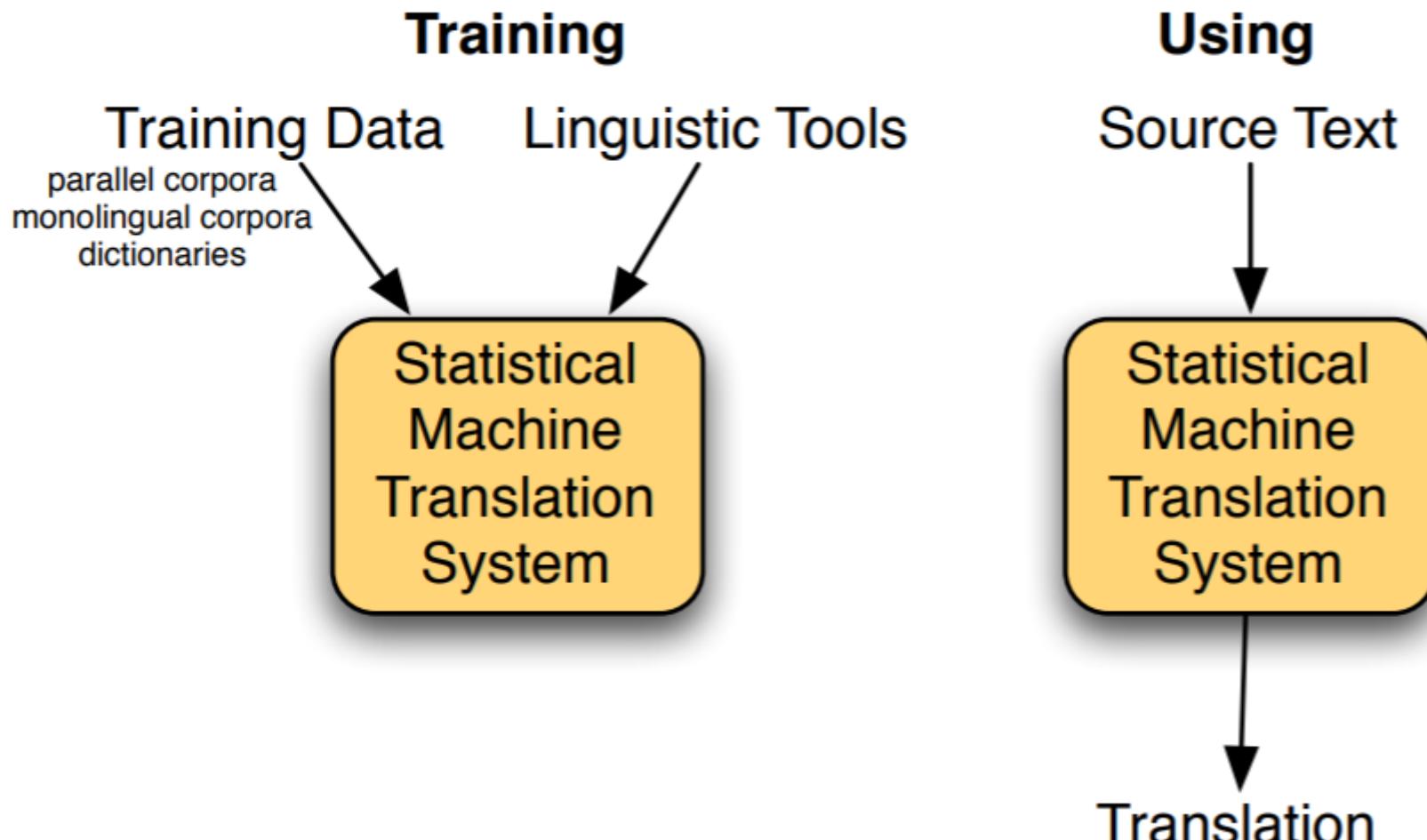


- Core idea: Learn a *probabilistic model* from data
- Suppose we're translating French → English.
- We want to find *best English sentence* y , given French sentence x
$$\operatorname{argmax}_y P(y|x)$$
- Use Bayes Rule to break this down into *two components* to be learnt separately:

$$= \operatorname{argmax}_y P(x|y)P(y)$$



Statistical Machine Translation



Parallel Corpora

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .
3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .
4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .
5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .
6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok
7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .
8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .
9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .
12b. wat nnat forat arrat vat gat .

Translation challenge: **farok crrrok hihok yorok clok kantok ok-yurp**

Statistical Translation: Learning from Data



- What is the best translation?

Sicherheit → security 14,516

Sicherheit → safety 10,015

Sicherheit → certainty 334

- Phrasal rules

Sicherheitspolitik → security policy 1580

Sicherheitspolitik → safety policy 13

Sicherheitspolitik → certainty policy 0

Lebensmittelsicherheit → food security 51

Lebensmittelsicherheit → food safety 1084

Lebensmittelsicherheit → food certainty 0

Rechtssicherheit → legal security 156

Rechtssicherheit → legal safety 5

Rechtssicherheit → legal certainty 723

Parallel Corpora

Look at a parallel corpus (German text along with English translation)

Translation of <i>Haus</i>	Count
house	8,000
building	1,600
home	200
household	150
shell	50

Maximum likelihood estimation

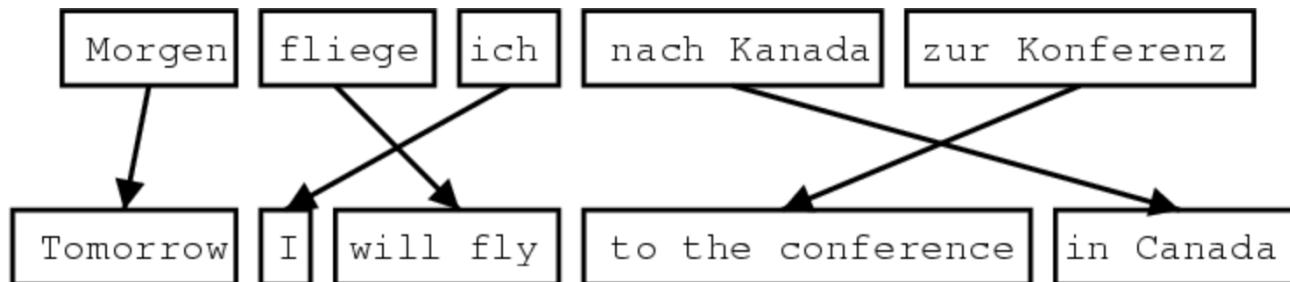
$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \text{house}, \\ 0.16 & \text{if } e = \text{building}, \\ 0.02 & \text{if } e = \text{home}, \\ 0.015 & \text{if } e = \text{household}, \\ 0.005 & \text{if } e = \text{shell}. \end{cases}$$

Learning alignment for SMT

- Question: How to learn translation $P(x|y)$ from model the parallel corpus?
- Break it down further: Introduce latent a variable into the model:

$$P(x, a|y)$$

where a is the **alignment**, i.e. word-level correspondence between source sentence x and target sentence y



Learning alignment for SMT



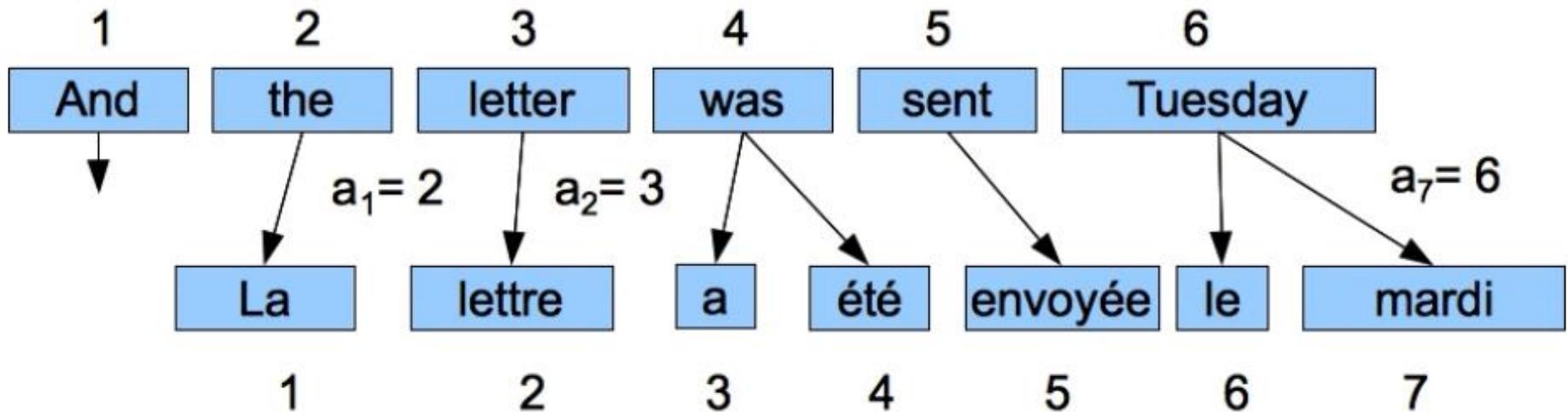
- We learn $P(x, a|y)$ as a combination of many factors, including:
 - Probability of particular words aligning (also depends on position in sent)
 - Probability of particular words having particular fertility (number of corresponding words)
 - etc.
- Alignments a are **latent variables**: They aren't explicitly specified in the data!
 - Require the use of special learning algos (like Expectation- Maximization) for learning the parameters of distributions with latent variables

Word Alignment

- Directly constructing phrase alignments is difficult, so rely on first constructing word alignments.
- Can learn to align from supervised word alignments, but human-aligned bitexts are rare and expensive to construct.
- Typically use an unsupervised EM-based approach to compute a word alignment from unannotated parallel corpus.

Word Alignment

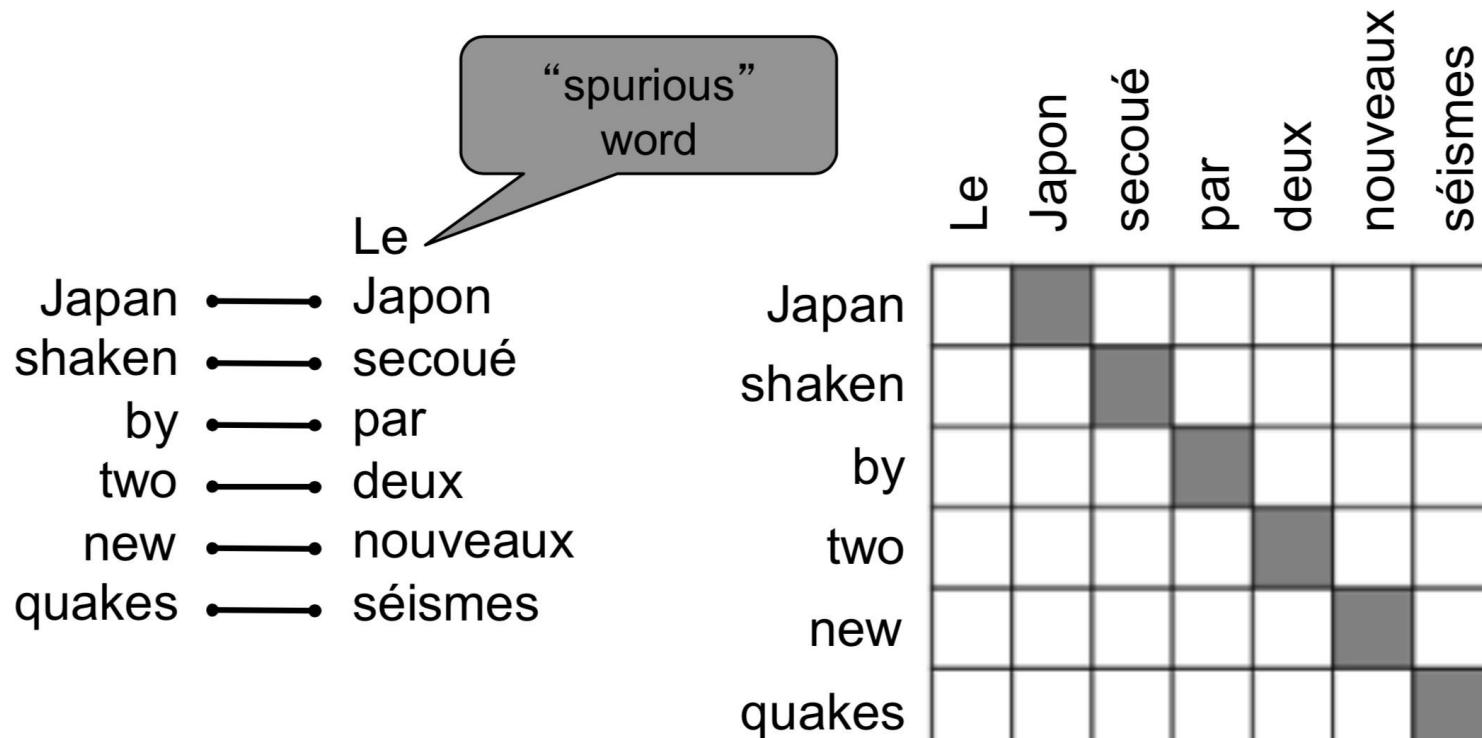
- Shows mapping between words in one language and the other.



What is alignment?

Alignment is the correspondence between particular words in the translated sentence pair.

- Typological differences between languages lead to complicated alignments!
- Note: Some words have no counterpart

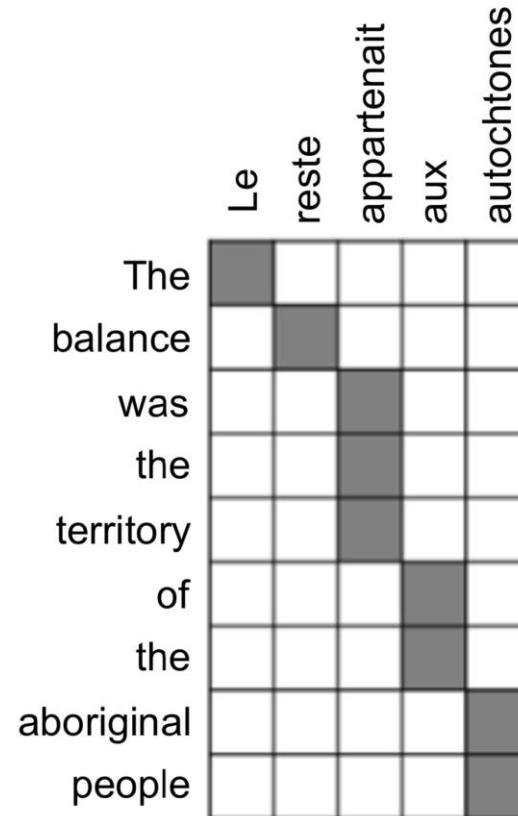
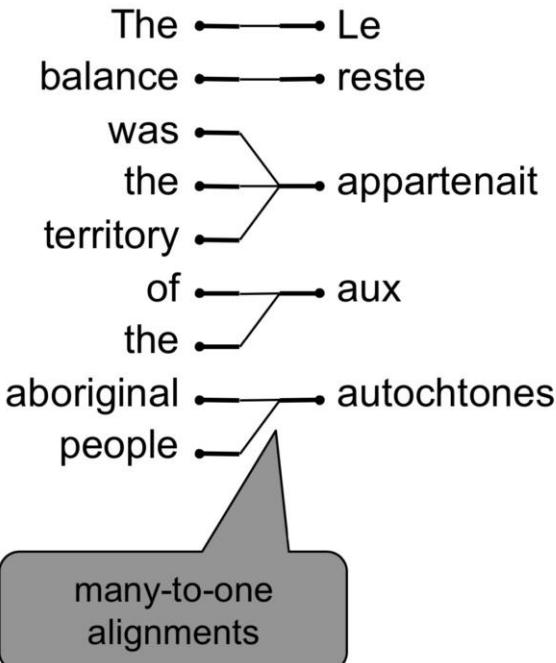


Examples from: “The Mathematics of Statistical Machine Translation: Parameter Estimation”, Brown et al, 1993.
<http://www.aclweb.org/anthology/J93-2003>

Alignment is complex



Alignment can be *many-to-one*



Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993.
<http://www.aclweb.org/anthology/J93-2003>

Alignment is complex

Alignment can be one-to-many

We call this a
fertile word

And
the
program
has
been
implemented

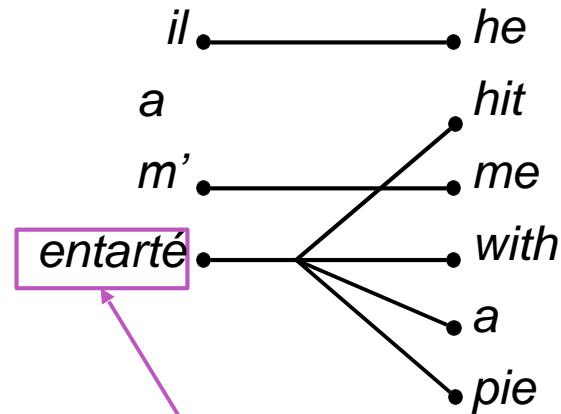
Le
programme
a
été
mis
en
application

one-to-many
alignment

Le	programme								
a		été							
mis									
en									
application									
And									
the									
program									
has									
been									
implemented									

Alignment is complex

Some words are very fertile!



This word has no single-word equivalent in English

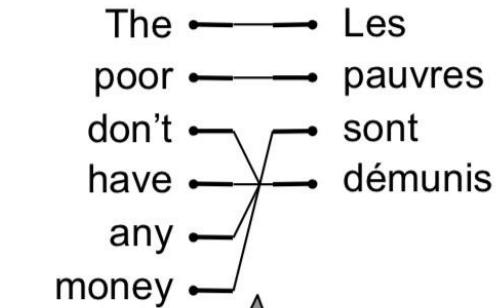
he hit me with a pie

il					
a					
m'					
entarté					

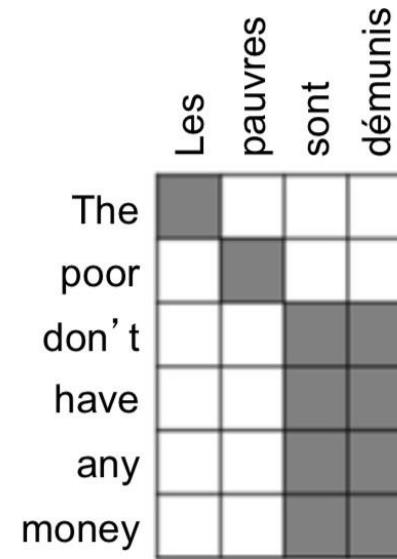


Alignment is complex

Alignment can be many-to-many (phrase-level)



many-to-many
alignment



The diagram shows phrase alignments between two sentences:

Les	pauvres	sont	démunis
The	poor	don't	have
any			
money			

phrase
alignment

Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993.
<http://www.aclweb.org/anthology/J93-2003>

Noisy Channel Model

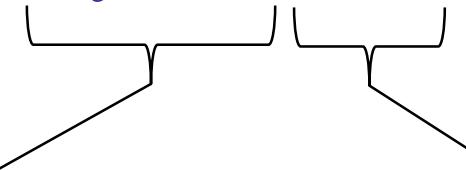
- Based on analogy to information-theoretic model used to decode messages transmitted via a communication channel that adds errors.
- Assume that source sentence was generated by a “noisy” transformation of some target language sentence and then use Bayesian analysis to recover the most likely target sentence that generated it.

Translate foreign language sentence $f = f_1, f_2, \dots, f_m$ to an English sentence $\hat{e} = e_1, e_2, \dots, e_I$ that maximizes $P(e | f)$

Bayesian Analysis of Noisy Channel

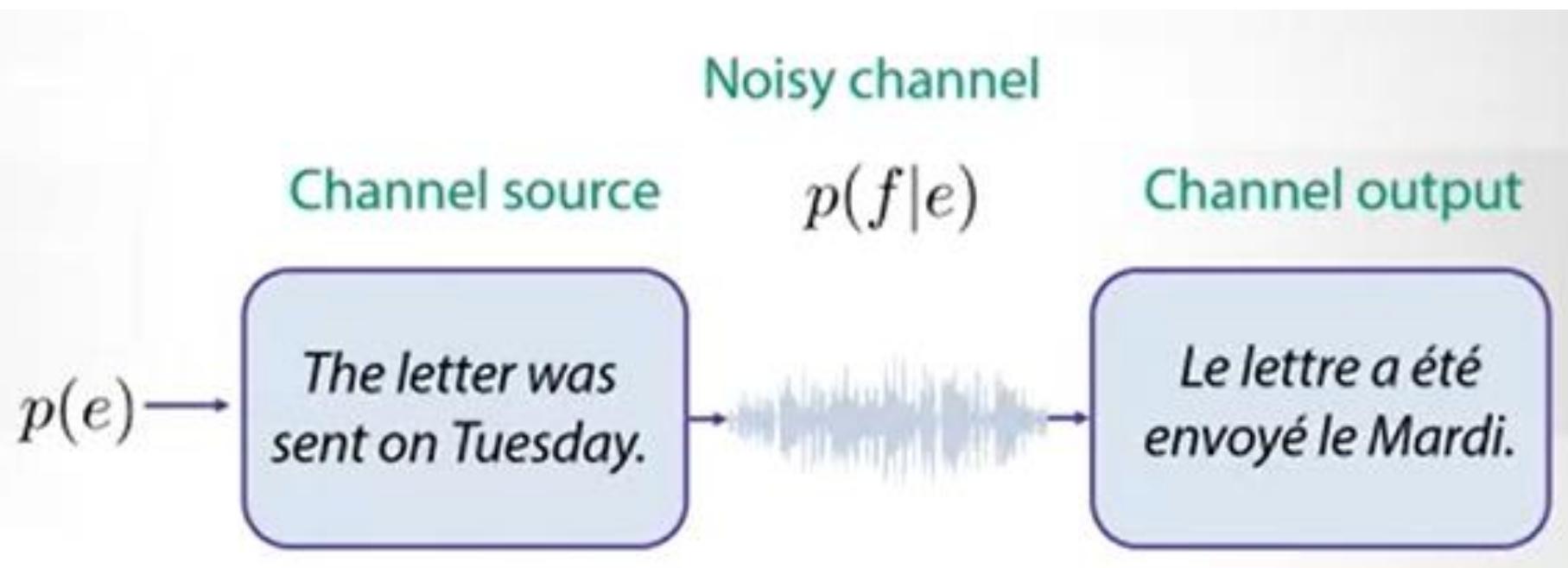


$$\begin{aligned}\hat{e} &= \operatorname{argmax} P(e | f) \\ &= \operatorname{argmax} P(f | e)P(e) / P(f) \\ &= \operatorname{argmax} P(f | e)P(e)\end{aligned}$$



A *decoder* determines the most probable translation \hat{e} given f

Noisy Channel: Translation Model



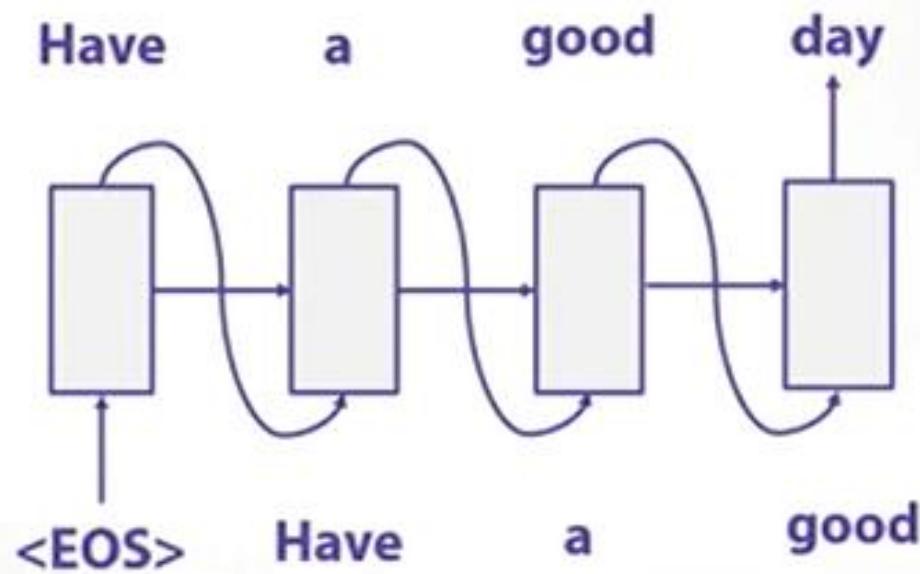
Language Model

- Use a standard n -gram language model for $P(e)$.
- Can be trained on a large, unsupervised mono-lingual corpus for the target language e .
- Could use a more sophisticated PCFG language model to capture long-distance dependencies.
- Terabytes of web data have been used to build a large 5-gram model of English.

Language Model

$$p(\mathbf{e}) = p(e_1)p(e_2|e_1) \dots p(e_k|e_1 \dots e_{k-1})$$

N-gram models or neural networks:



Language model

- What is most fluent?

a problem for translation 13,000
a problem of translation 61,600
a problem in translation 81,700

- Hits on Google

Translation Model

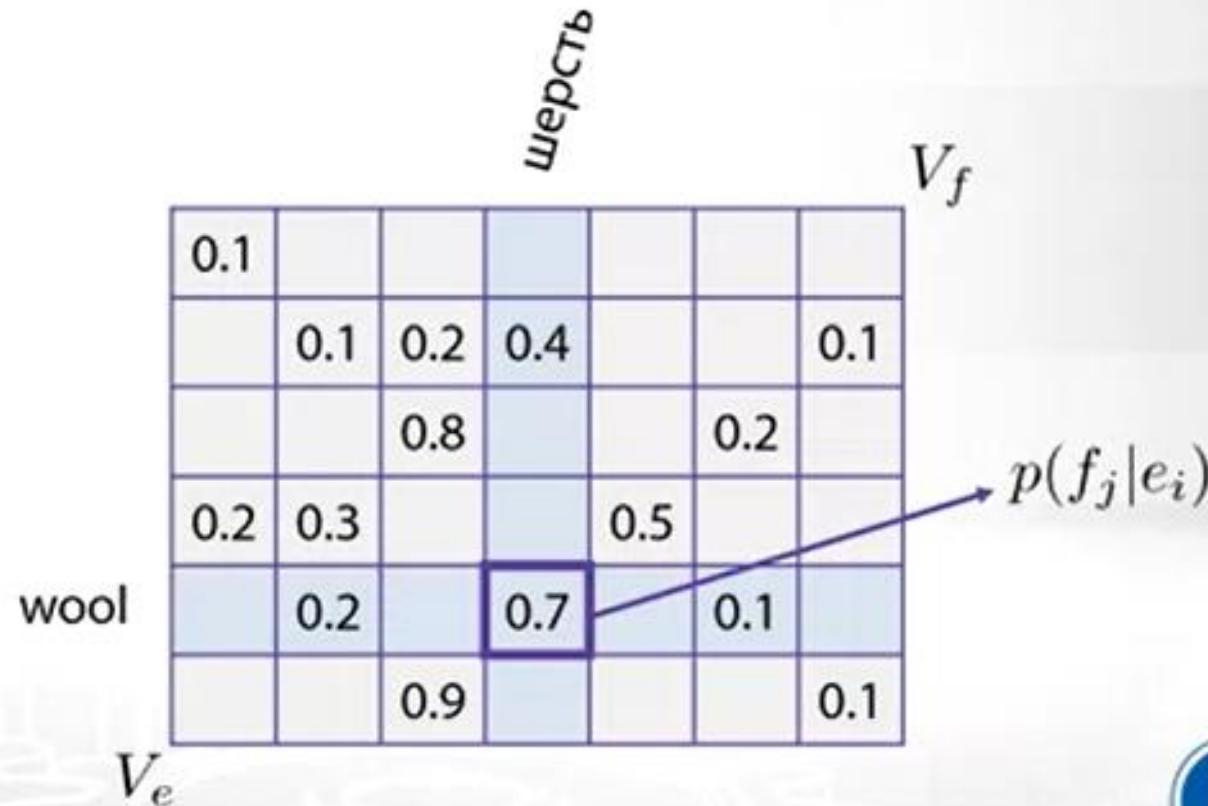
$$p(f|e) = p(f_1, f_2, \dots, f_J | e_1, e_2, \dots, e_I)$$

f (Foreign): Крику много, а шерсти мало.

e (English): Great cry and little wool.

Translation Model

We could learn translation probabilities for separate words:



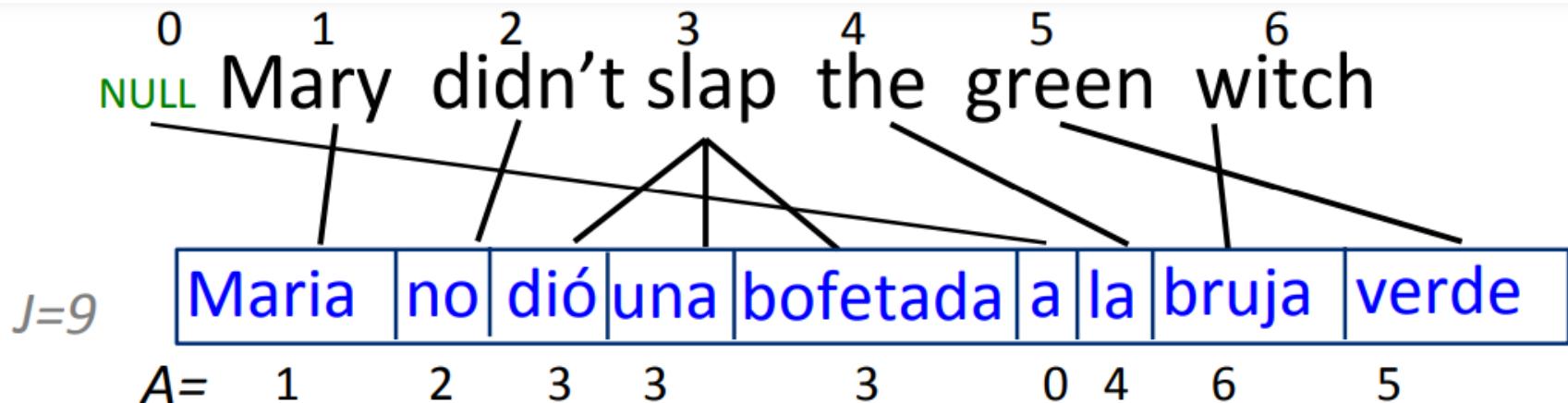
IBM Model 1

- First model proposed in seminal paper by Brown *et al.* in 1993 as part of CANDIDE, the first complete SMT system.

Simple generative model to produce F given $E=e_1, e_2, \dots e_I$

- Choose J , the number of words in F : $F=f_1, f_2, \dots f_J$
- Choose a 1-to-many alignment $A=a_1, a_2, \dots a_J$
- For each position in F , generate a word f_j from the aligned word in E : e_{a_j}

IBM Model 1



1. Choose J , the number of words in F : $F=f_1, f_2, \dots, f_J$
2. Choose a 1-to-many alignment $A=a_1, a_2, \dots, a_J$
3. For each position in F , generate a word f_j from the aligned word in E : e_{a_j}

IBM Model 1

- Let
 - e_{a_j} : the English word assigned to Spanish word f_j
 - $t(f_x, e_y)$: probability of translating e_y as f_x
- If we knew E, the alignment A, and J, then:

$$P(F | E, A) = \prod_{j=1}^J t(f_j, e_{a_j})$$

- The probability of the Spanish sentence if we knew the English source, the alignment, and J

IBM Model 1

- A normalization factor, since there are $(I + 1)^J$ possible alignments:

$$P(A | E) = \frac{\varepsilon}{(I + 1)^J}$$

- The probability of an alignment given the English sentence.

IBM Model 1

$$P(A | E) = \frac{\varepsilon}{(I+1)^J} \quad P(F | E, A) = \prod_{j=1}^J t(f_j, e_{a_j})$$

The probability of generating F through a particular alignment:

$$P(F, A | E) = \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^J t(f_j, e_{a_j})$$

To get $P(F | E)$, we sum over all alignments:

$$P(F | E) = \sum_A P(F, A | E) = \sum_A \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^J t(f_j, e_{a_j})$$

IBM Model 1

- Goal is to find the most probable alignment given a parameterized model.

$$\begin{aligned}
 \hat{A} &= \operatorname{argmax}_A P(F, A | E) \\
 &= \operatorname{argmax}_A \frac{P(J | E)}{(I+1)^J} \prod_{j=1}^J t(f_j, e_{a_j}) \\
 &= \operatorname{argmax}_A \prod_{j=1}^J t(f_j, e_{a_j})
 \end{aligned}$$

Since translation choice for each position j is independent, the product is maximized by maximizing each term:

$$a_j = \operatorname{argmax}_{0 \leq i \leq I} t(f_j, e_i) \quad 1 \leq j \leq J$$

IBM Model 1 : EM Algorithm

Incomplete data – if we had complete data, would could estimate model – if we had model, we could fill in the gaps in the data

Expectation Maximization (EM) in a nutshell

1. initialize model parameters (e.g. uniform)
2. assign probabilities to the missing data
3. estimate model parameters from completed data
4. iterate steps 2–3 until convergence

Sketch of EM Algorithm for Word Alignment

Randomly set model parameters.

(making sure they represent legal distributions)

Until converge (i.e. parameters no longer change) do:

E Step: Compute the probability of all possible alignments of the training data using the current model.

M Step: Use these alignment probability estimates to re-estimate values for all of the parameters.

Note: Use dynamic programming (as in Baum-Welch) to avoid explicitly enumerating all possible alignments

IBM Model 1

- Simplified version of Model 1
(No NULL word, and subset of alignments: ignore alignments for which English word aligns with no foreign word)

- E-step

$$P(A, F | E) = \prod_{j=1}^J t(f_j | e_{a_j})$$

(ignoring a constant here)

- Normalize to get probability of an alignment:

$$P(A | E, F) = \frac{P(A, F | E)}{\sum_A P(A, F | E)} = \frac{\prod_{j=1}^J t(f_j | e_{a_j})}{\sum_A \prod_{j=1}^J t(f_j | e_{a_j})}$$

Sample EM Trace for Alignment

(IBM Model 1 with no NULL Generation)

*Training
Corpus*

green house
casa verde

the house
la casa

*Translation
Probabilities*

	<i>verde</i>	<i>casa</i>	<i>la</i>
<i>green</i>	1/3	1/3	1/3
<i>house</i>	1/3	1/3	1/3
<i>the</i>	1/3	1/3	1/3

*Assume uniform
initial probabilities*

*Compute
Alignment
Probabilities*

$P(a, f | e)$

*Normalize
to get
 $P(a | f, e)$*

green house
casa verde

green house
~~*casa verde*~~

the house
la casa

the house
~~*la casa*~~

$$1/3 \times 1/3 = 1/9$$

$$\frac{1/9}{2/9} = \frac{1}{2}$$

$$\frac{1/9}{2/9} = \frac{1}{2}$$

$$\frac{1/9}{2/9} = \frac{1}{2}$$

$$\frac{1/9}{2/9} = \frac{1}{2}$$

Example cont.

green house
casa verde

1/2

~~*green house*~~
~~*casa verde*~~

1/2

the house
la casa

1/2

~~*the house*~~
~~*la casa*~~

1/2

Compute
weighted
translation
counts

	<i>verde</i>	<i>casa</i>	<i>la</i>
<i>green</i>	1/2	1/2	0
<i>house</i>	1/2	1/2 + 1/2	1/2
<i>the</i>	0	1/2	1/2

Normalize
rows to sum
to one to
estimate $P(f | e)$

	<i>verde</i>	<i>casa</i>	<i>la</i>
<i>green</i>	1/2	1/2	0
<i>house</i>	1/4	1/2	1/4
<i>the</i>	0	1/2	1/2

Example cont.

*Translation
Probabilities*

	<i>verde</i>	<i>casa</i>	<i>la</i>
<i>green</i>	1/2	1/2	0
<i>house</i>	1/4	1/2	1/4
<i>the</i>	0	1/2	1/2

*Recompute
Alignment
Probabilities
 $P(a, f / e)$*

<i>green house</i>	<i>green house</i>	<i>the house</i>	<i>the house</i>
<i>casa verde</i>	<i>casa verde</i>	<i>la casa</i>	<i>la casa</i>
$1/2 \times 1/4 = 1/8$	$1/2 \times 1/2 = 1/4$	$1/2 \times 1/2 = 1/4$	$1/2 \times 1/4 = 1/8$

*Normalize
to get
 $P(a / f, e)$*

$$\frac{1/8}{3/8} = \frac{1}{3} \quad \frac{1/4}{3/8} = \frac{2}{3} \quad \frac{1/4}{3/8} = \frac{2}{3} \quad \frac{1/8}{3/8} = \frac{1}{3}$$

*Continue EM iterations until translation
parameters converge*

HMM-Based Word Alignment

- IBM Model 1 assumes all alignments are equally likely and does not take into account ***locality***:
 - If two words appear together in one language, then their translations are likely to appear together in the result in the other language.
- An alternative model of word alignment based on an HMM model ***does*** account for locality by making longer jumps in switching from translating one word to another less likely.

HMM Model

- Assumes the hidden state is the specific word occurrence e_i in e currently being translated (i.e. there are l states, one for each word in e).
- Assumes the observations from these hidden states are the possible translations f_j of e_i .
- Generation of f from e then consists of moving to the initial e word to be translated, generating a translation, moving to the next word to be translated, and so on.

Sample HMM Generation



Mary didn't slap the green witch.

1 2 3 4 5 6

Maria

```
graph TD; Mary[1 Mary] --> Maria[2 Maria]; didn't[2 didn't] --> Maria; slap[3 slap] --> Maria; the[4 the] --> Maria; green[5 green] --> Maria; witch[6 witch.] --> Maria;
```

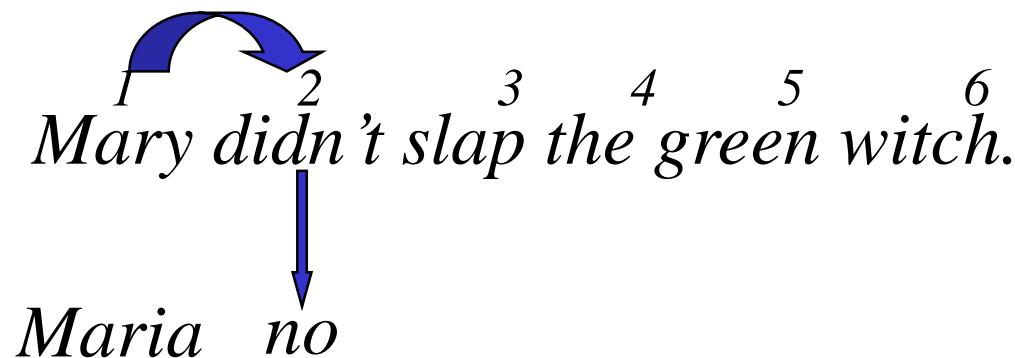
Sample HMM Generation



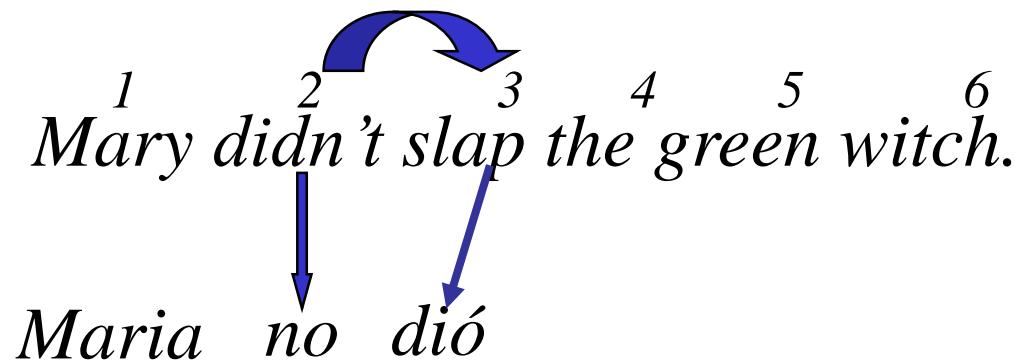
Mary didn't slap the green witch.

1 2 3 4 5 6

Maria no



Sample HMM Generation



Sample HMM Generation



Mary didn't slap the green witch.

Maria no dió una



Sample HMM Generation



Mary didn't slap the green witch.

Maria no dió una bofetada



Sample HMM Generation



Mary didn't slap the green witch.

Maria no dió una bofetada a

Sample HMM Generation



Mary didn't slap the green witch.

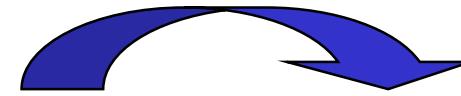
Maria no dió una bofetada a la



Sample HMM Generation



Mary didn't slap the green witch.

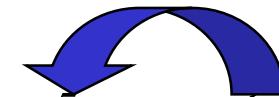


Maria no dió una bofetada a la bruja

Sample HMM Generation



Mary didn't slap the green witch.



Maria no dió una bofetada a la bruja verde.

Sample HMM Generation



Mary ¹ didn't ² slap ³ the ⁴ green ⁵ witch. ⁶

Maria no dió una bofetada a la bruja verde.

HMM Parameters

- Transition and observation parameters of states for HMMs for all possible source sentences are “tied” to reduce the number of free parameters that have to be estimated.
- **Observation probabilities:** $b_j(f_i) = P(f_i | e_j)$ the same for all states representing an occurrence of the same English word.
- **State transition probabilities:** $a_{ij} = s(j-i)$ the same for all transitions that involve the same *jump width* (and direction).

Computing $P(F | E)$ in the HMM Model

- Given the observation and state-transition probabilities, $P(f | e)$ (observation likelihood) can be computed using the standard ***forward algorithm*** for HMMs.

Decoding for the HMM Model



- Use the standard ***Viterbi algorithm*** to efficiently compute the most likely alignment (i.e. most likely state sequence).

Training Word Alignment Models



- Both the IBM model 1 and HMM model can be trained on a parallel corpus to set the required parameters.
 - For supervised (hand-aligned) training data, parameters can be estimated directly using frequency counts.
 - For unsupervised training data, EM can be used to estimate parameters, e.g. Baum-Welch for the HMM model.
-

Evaluating MT

- Human subjective evaluation is the best but is time-consuming and expensive.
- Automated evaluation comparing the output to multiple human reference translations is cheaper and correlates with human judgements.

Translation Quality

- Achieving literary quality translation is very difficult.
 - Existing MT systems can generate rough translations that frequently at least convey the gist of a document.
 - High quality translations possible when specialized to narrow domains, e.g. weather forecasts.
 - Some MT systems used in ***computer-aided translation*** in which a bilingual human post-edits the output to produce more readable accurate translations.
 - Frequently used to aid ***localization*** of software interfaces and documentation to adapt them to other languages.
-

Human Evaluation of MT

- Ask humans to estimate MT output on several dimensions.
 - **Fluency:** Is the result grammatical, understandable, and readable in the target language.
 - **Fidelity/faithfulness:** Does the result correctly convey the information in the original source language.
 - **Adequacy/Acceptability:**
 - Human judgment on a fixed scale.
 - Bilingual judges given source and target language.
 - Monolingual judges given reference translation and MT result.
 - **Informativeness:** Monolingual judges must answer questions about the source sentence given only the MT translation (task-based evaluation).

Picking a Good Translation

- A good translation should be *faithful* and correctly convey the information and tone of the original source sentence.
- A good translation should also be *fluent*, grammatically well structured and readable in the target language.
- Final objective:

$$T_{best} = \underset{T \in \text{Target}}{\operatorname{argmax}} \text{ faithulne ss}(T, S) \text{ fluency}(T)$$

Faithfulness: $P(F | E)$

- *Spanish:*
 - *Maria no dió una bofetada a la bruja verde*
- *English candidate translations:*
 - *Mary didn't slap the green witch*
 - *Mary not give a slap to the witch green*
 - *The green witch didn't slap Mary*
 - *Mary slapped the green witch*
- *More faithful translations will be composed of phrases that are high probability translations*
 - *How often was “slapped” translated as “dió una bofetada” in a large **bitext** (parallel English–Spanish corpus)*
 - *We’ll need to align phrases and words to each other in bitext*

Issues in evaluation

Different word order conveying the same message

I was late for office due to traffic jam

The traffic jam was responsible for my delay to office

Traffic jam delayed me to office

Computer-Aided Translation Evaluation

Edit cost

Measure the number of changes that a human translator must make to correct the MT output.

- Number of words changed
- Amount of time taken to edit
- Number of keystrokes needed to edit

Automatic Evaluation of MT

- Collect one or more human *reference translations* of the source.
- Compare MT output to these reference translations.
- Score result based on similarity to the reference translations.
 - Precision
 - BLEU
 - NIST

Precision

Precision = $\frac{\text{No. of candidate translation words occurring in any reference translation}}{\text{Total no. of words in the candidate translation}}$

Candidate 1: the the the the the the the.

Candidate 2: the cat is .nat the on

Reference: The cat is on the mat.

- The precision for candidate 1 is 2/7 (28.5%)
- The Precision for candidate 2 is 1(100%).

BLEU-Bilingual Evaluation Understudy

- Closer a machine translation is to a professional human translation, the better it is
- Determine number of n -grams of various sizes that the MT output shares with the reference translations.
 - ***Reference translation is Human translation***
 - ***Candidate Translation is Machine translation***
- Compute a modified precision measure of the n -grams in MT result.

BLEU-Bilingual Evaluation Understudy



Step 1: For each *n-gram* in the candidate, we count $C(w)$ how many times it appears in the candidate. Calculate ***total*** count

Step 2: For each *n-gram* define $R(w)$ to be the largest number of times the *n-gram* appears in any of the references.

Step 3: ***Clip*** count: $\text{MIN}(R(w), C(w))$ where MIN is the minimum of the two values.

Step 4: Bleu Score= Clip count / Total

Bleu- Example 1

Reference

- R1: but thou shalt **love** thy neighbor as thyself
- R2: but have **love** for your neighbor as for yourself
- R3: but **love** your neighbors as you **love** yourself

Candidate

C: but **love** other **love** friend for **love** yourself

Step 1: Count each n-gram occurrence in **candidate**

C(but) =1

C(**love**) =3

C(Other) =1

C(friend) =1

C(For) =1

C(Yourself) = 1

Total=8

Bleu- Example 1

Step 2: Reference Counts

Maximum occurrence

R(but)=1

R(love)=2 [appears twice in R3]

R(other)=0

R(friend)=0

R(for)=2 [appears twice in R2]

R(yourself)=1

Step 3: Clip count

minimum of candidate and reference

$\text{MIN}(C(\text{but}), R(\text{but})) = \text{MIN}(1, 1) = 1$

$\text{MIN}(C(\text{love}), R(\text{love})) = \text{MIN}(3, 2) = 2$

$\text{MIN}(C(\text{other}), R(\text{other})) = \text{MIN}(1, 0) = 0$

$\text{MIN}(C(\text{friend}), R(\text{friend})) = \text{MIN}(1, 0) = 0$

$\text{MIN}(C(\text{for}), R(\text{for})) = \text{MIN}(1, 2) = 1$

$\text{MIN}(C(\text{yourself}), R(\text{yourself})) = \text{MIN}(1, 1) = 1$

Total clip count=5 summation of clip count

Step 4:

Bleu Score:= Total clip count / Total = 5/8

BLEU Example 2

Candidate 1: Mary no slap the witch green.

*Ref 1: Mary did not **slap** the green witch.*

Ref 2: Mary did not smack the green witch.

Ref 3: Mary did not hit a green sorceress.

Candidate 1 Unigram Precision: 5/6

BLEU Example 2

*Candidate 1: Mary no **slap the** witch green.*

*Ref 1: Mary did not **slap the** green witch.*

Ref 2: Mary did not smack the green witch.

Ref 3: Mary did not hit a green sorceress.

Candidate 1 Bigram Precision: 1/5

BLEU Example 3

Candidate 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.

Ref 2: Mary did not smack the green witch.

Ref 3: Mary did not hit a green sorceress.

Clip match count of each n-gram to maximum count of the n-gram in any single reference translation

Candidate 2 Unigram Precision: 7/10

BLEU Example 3

Candidate 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.

Ref 2: Mary did not smack the green witch.

Ref 3: Mary did not hit a green sorceress.

Clip match count of each n-gram to maximum count of the n-gram in any single reference translation

Candidate 2 Bigram Precision: 4/9

Modified N-Gram Precision

Average n -gram precision over all n -grams up to size N (typically 3 or 4) using geometric mean.

$$p_n = \frac{\sum_{C \in corpus} \sum_{n\text{-gram} \in C} \text{count}_{clip}(n\text{-gram})}{\sum_{C \in corpus} \sum_{n\text{-gram} \in C} \text{count}(n\text{-gram})}$$

$$p = \sqrt[N]{\prod_{n=1}^N p_n}$$

$$\text{Cand 1: } p = \sqrt[2]{\frac{5}{6} \frac{1}{5}} = 0.408$$

$$\text{Cand 2: } p = \sqrt[2]{\frac{7}{10} \frac{4}{9}} = 0.558$$

Brevity Penalty

- Use a penalty for translations that are shorter than the reference translations.
- Define effective reference length, r , for each sentence as the length of the reference sentence with the largest number of n -gram matches. Let c be the candidate sentence length.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

BLEU Score

Final BLEU Score: $\text{BLEU} = BP \times p$

Cand 1: Mary no slap the witch green.

Best Ref: Mary did not slap the green witch.

$$c = 6, \quad r = 7, \quad BP = e^{(1-7/6)} = 0.846$$

$$\text{BLEU} = 0.846 \times 0.408 = 0.345$$

Cand 2: Mary did not give a smack to a green witch.

Best Ref: Mary did not smack the green witch.

$$c = 10, \quad r = 7, \quad BP = 1$$

$$\text{BLEU} = 1 \times 0.558 = 0.558$$

BLEU Score Issues

- BLEU has been shown to correlate with human evaluation when comparing outputs from different SMT systems.
 - However, it does not correlate with human judgments when comparing SMT systems with manually developed MT (Systran) or MT with human translations.
 - Other MT evaluation metrics have been proposed that claim to overcome some of the limitations of BLEU.
-

NIST (National Institute of Standards and Technology)



It is based on the BLEU metric

BLEU simply calculates n-gram precision adding equal weight to each one

NIST also calculates how informative a particular n-gram is. That is to say when a correct n-gram is found, the rarer that n-gram is, the more weight it will be given.^[1]

For example, if the bigram "on the" is correctly matched, it will receive lower weight than the correct matching of bigram "interesting calculations", as this is less likely to occur.

References

- <https://www.coursera.org/lecture/language-processing/introduction-to-machine-translation-nv7Cr>
 - <https://www.translatefx.com/blog/what-is-neural-machine-translation-engine-how-does-it-work?lang=en>
 - <https://omniscien.com/faq/different-types-of-machine-translation/>
 - <https://arxiv.org/abs/2007.07691>
 - <http://mt-class.org/jhu/>
 - <https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1>
 - <https://medium.com/@ageitgey/build-your-own-google-translate-quality-machine-translation-system-d7dc274bd476>
 - <https://www.youtube.com/watch?v=AlpXjFwVdIE>
 - Dataset: <http://www.manythings.org/anki/>
-



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 10: Neural Machine translation

Date – 18th February 2024

Time – 1.40 pm to 3.40 pm

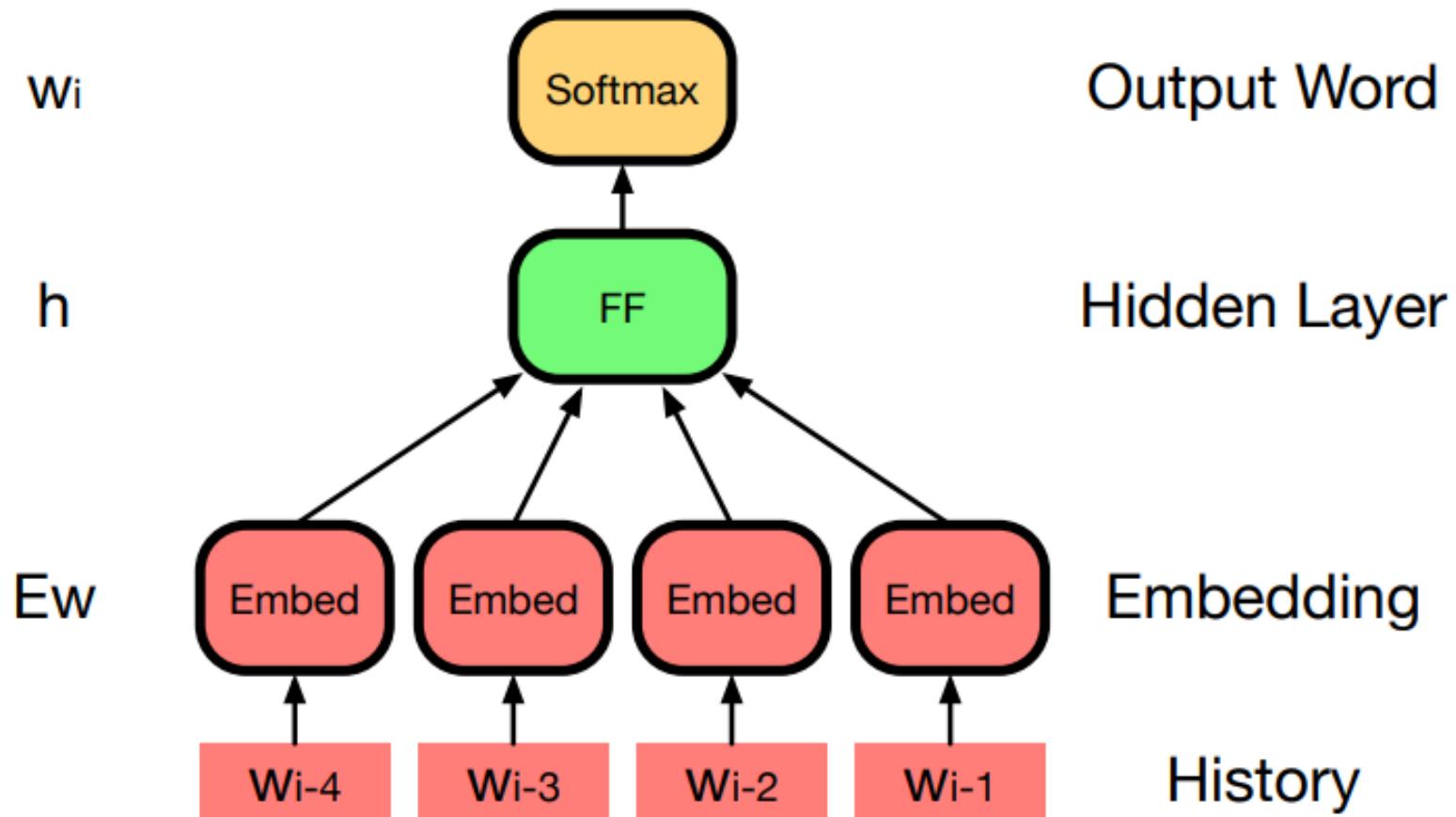
These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philipp Koehn**, Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

Agenda

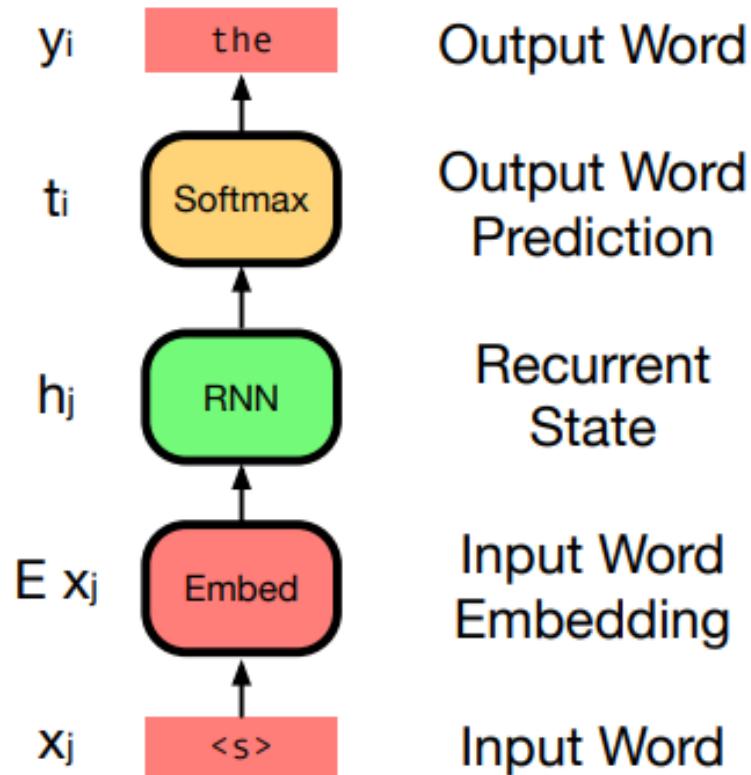
Neural Machine translation

- Introduction
- Recurrent Neural Translation Models
- Encoder Decoder Translation Models
- Neural translation model with attention
- Training Neural Models
- Deeper Models
- Demo

Feed Forward Neural Language Model

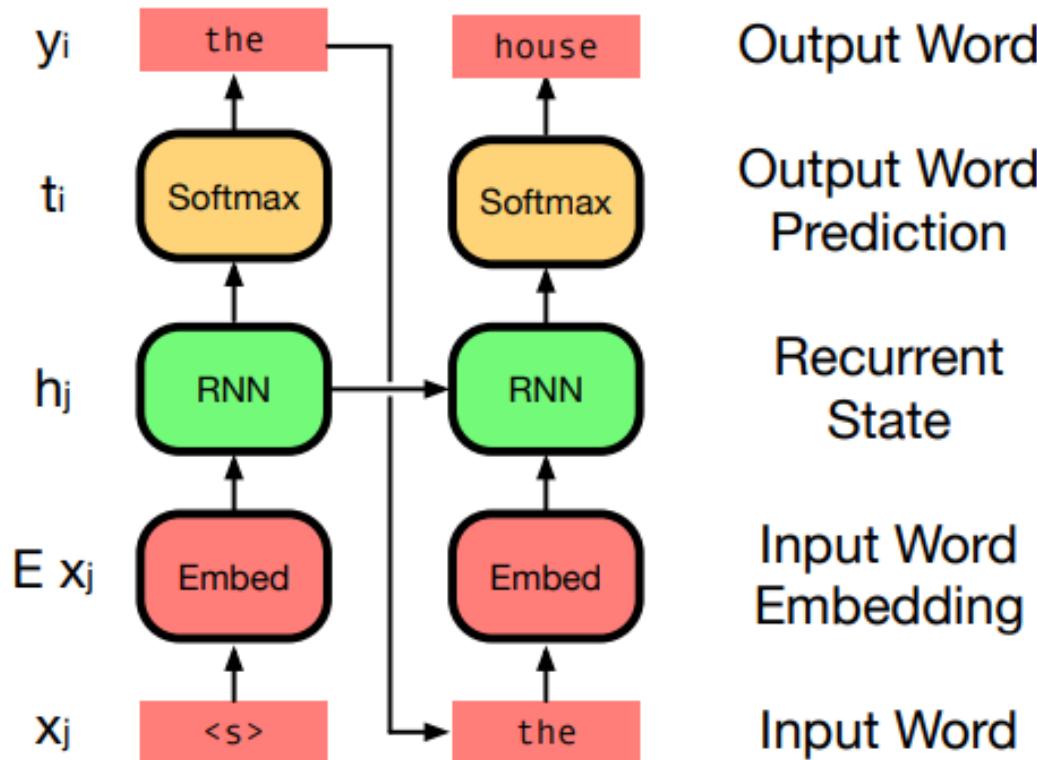


Recurrent Neural Language Model



Predict the first word of a sentence

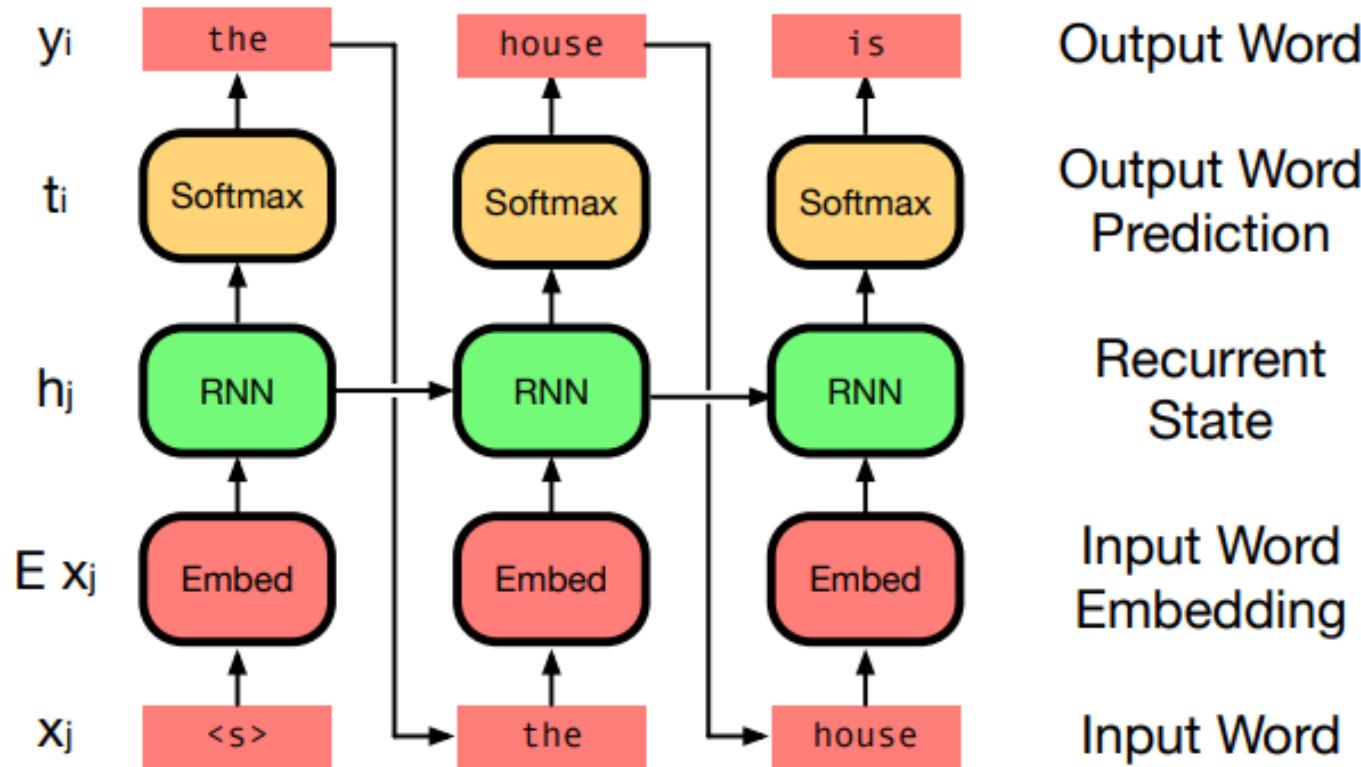
Recurrent Neural Language Model



Predict the second word of a sentence

Re-use hidden state from first word prediction

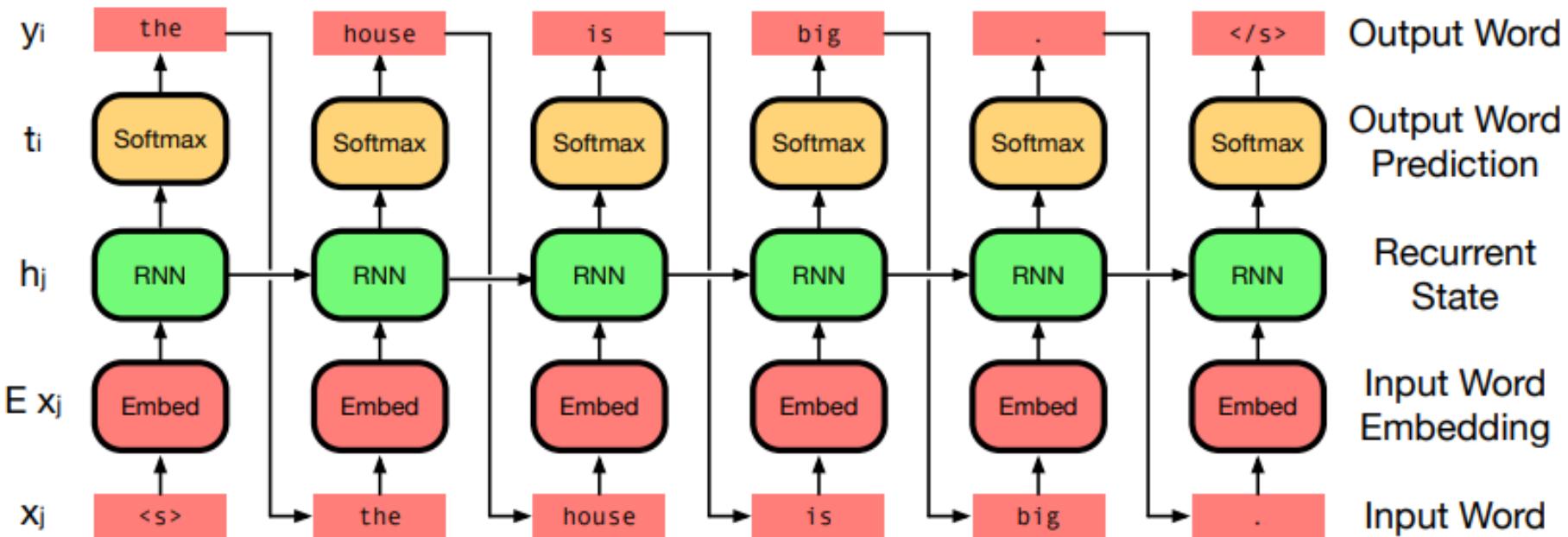
Recurrent Neural Language Model



Predict the third word of a sentence

... and so on

Recurrent Neural Language Model



What is Neural Machine Translation?

- Neural Machine Translation (NMT) is a way to do Machine Translation with a single neural network
- The neural network architecture is called sequence-to-sequence (aka seq2seq) and it involves two RNN

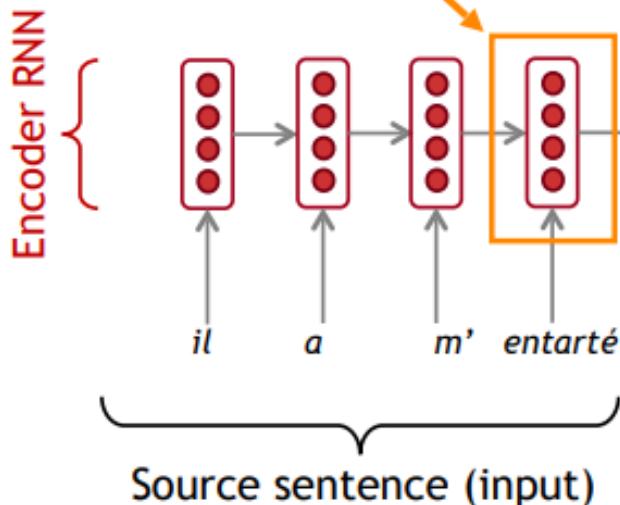
NN Machine translation?

- Neural Machine Translation is the task of converting a sequence of words from a source language, like English, to a sequence of words to a target language like Hindi or Spanish using deep neural networks.
- RNN's are neural networks with loops to persist information. They perform the same task for every element in the sequence and the output elements are dependent on previous elements or states.
- RNN work in two phases : Encoder and Decoder
- Uses algorithm “Teachers forcing” algorithm trains decoder by supplying actual output of the previous timestamp instead of the predicted output from the previous time as inputs during training.
- Embedding provides a dense representation of words and their relative meanings.

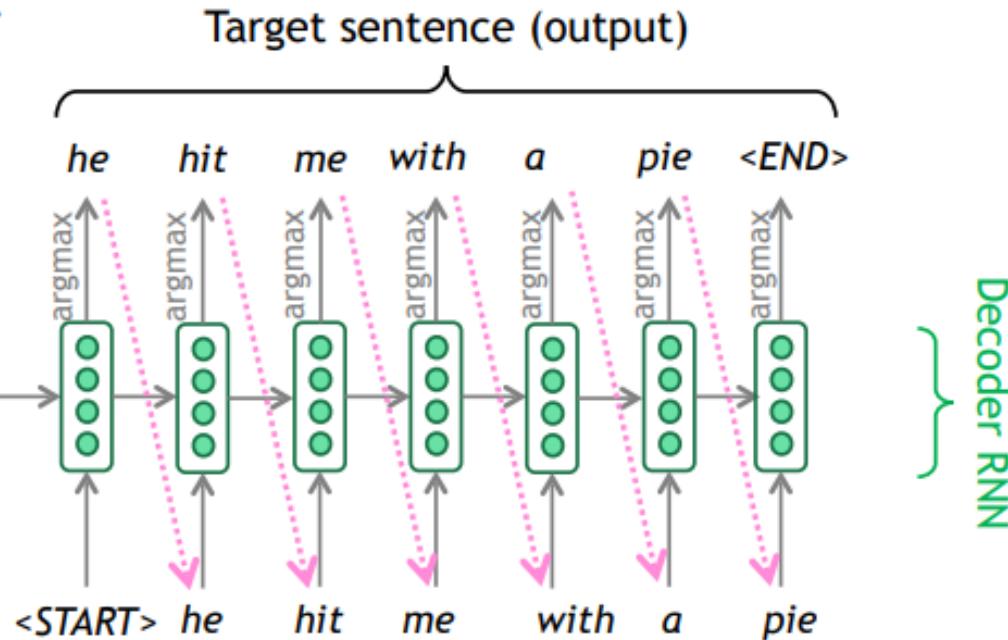
Neural Machine Translation (NMT)

The sequence-to-sequence model

Encoding of the source sentence.
Provides initial hidden state
for Decoder RNN.



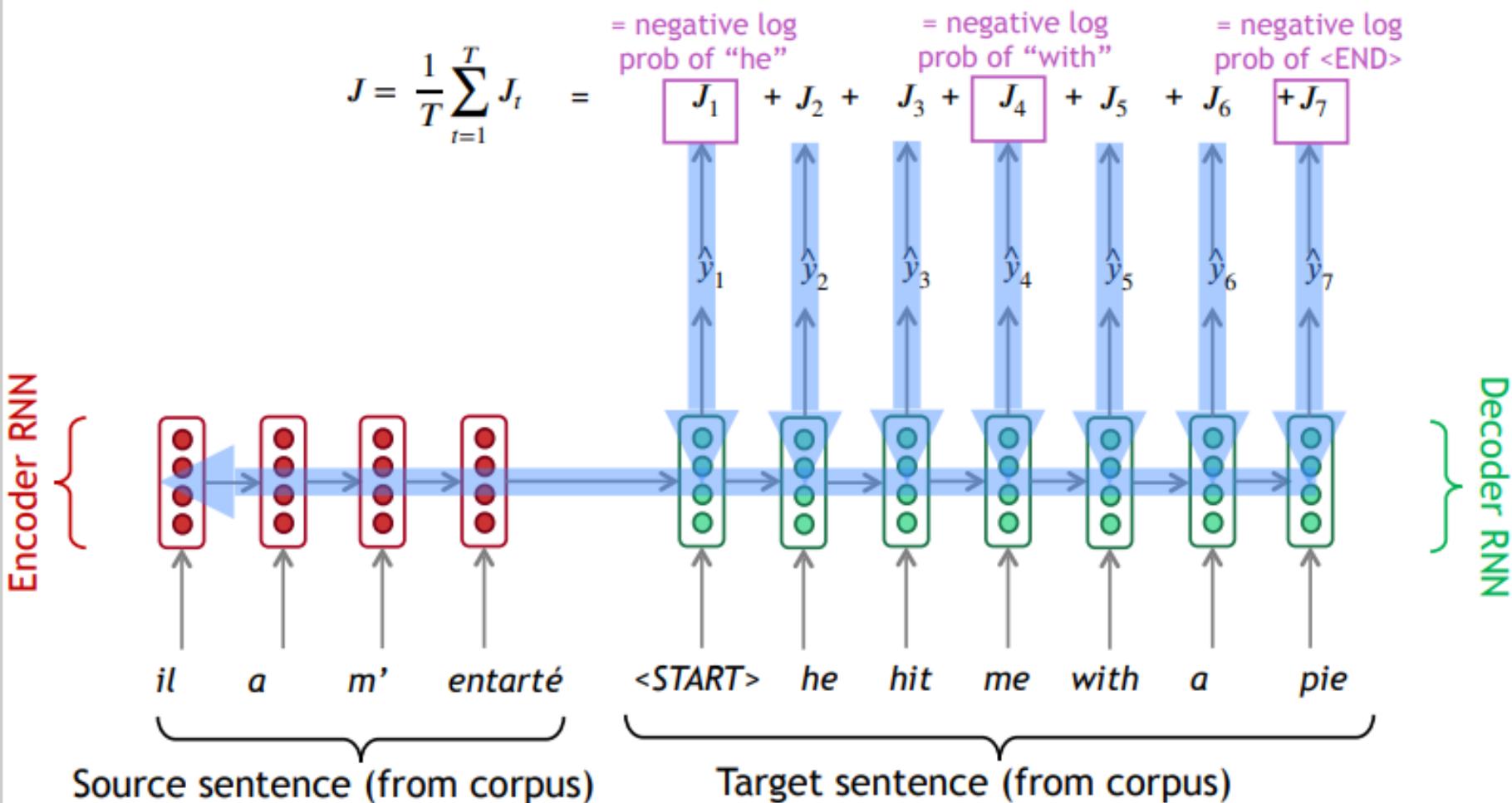
Encoder RNN produces
an encoding of the
source sentence.



Decoder RNN is a Language Model that generates target sentence, conditioned on encoding.

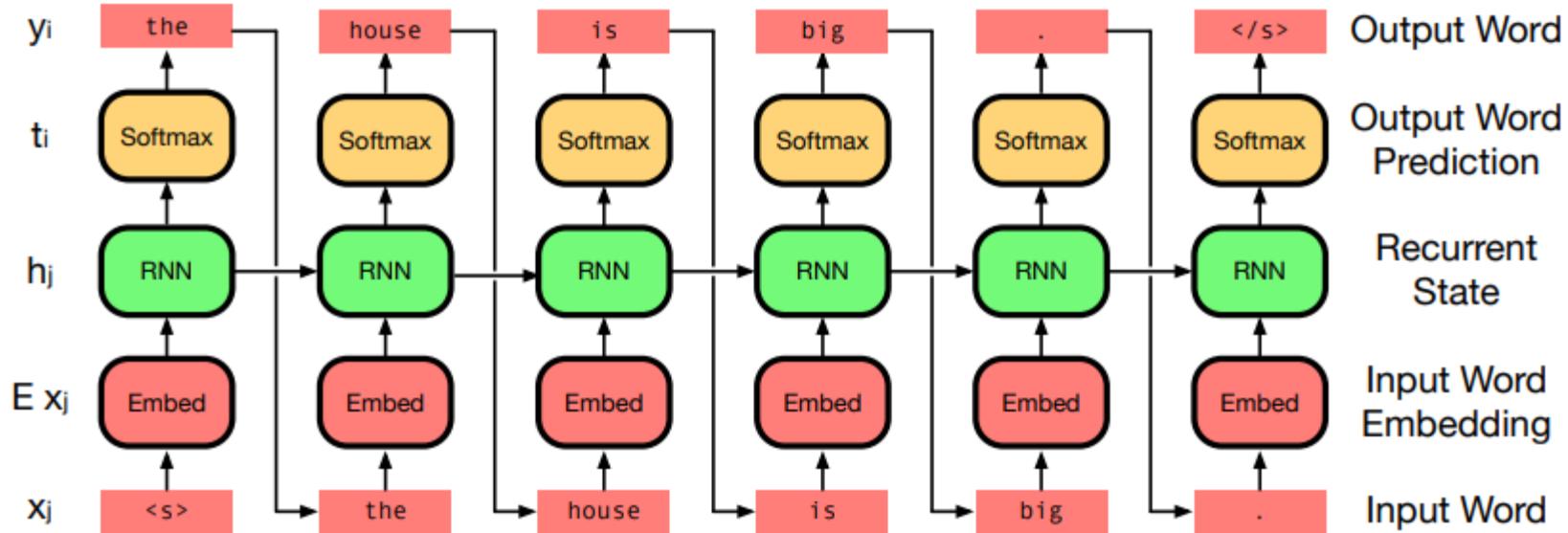
Note: This diagram shows test time behavior:
decoder output is fed in ➤ as next step's input

Training NMT



Neural translation model

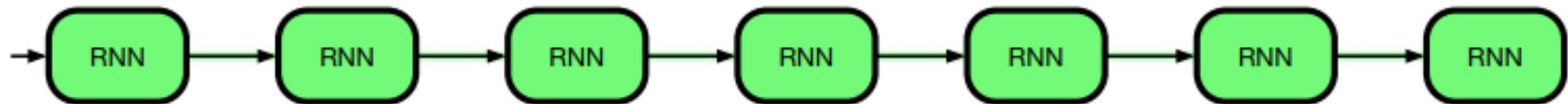
Input Encoding



- Inspiration: recurrent neural network language model on the input side

Hidden Language Model States

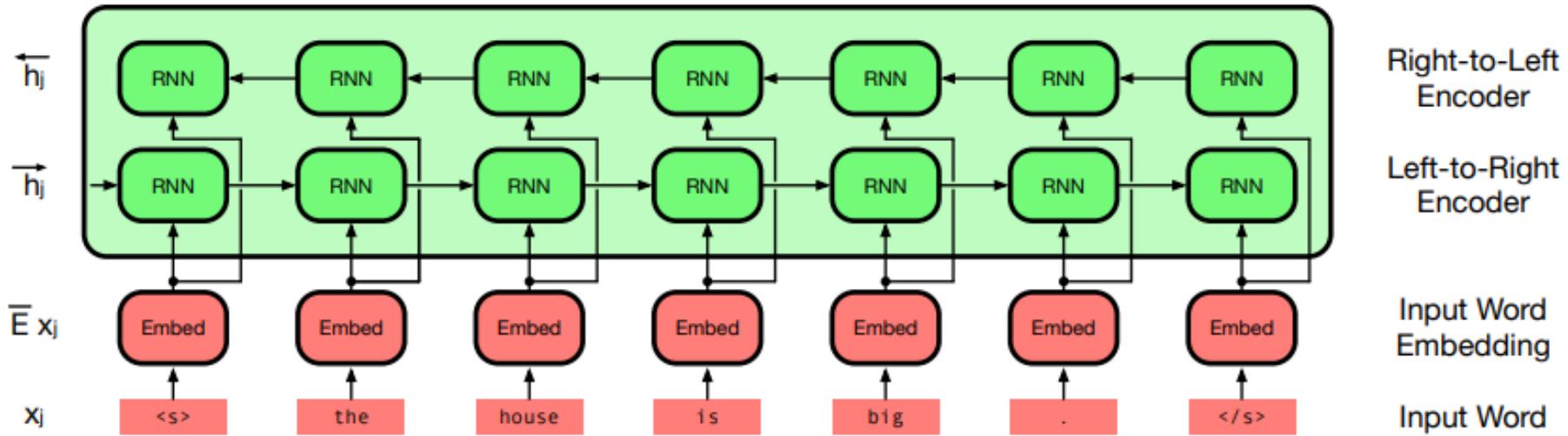
- This gives us the hidden states



- These encode left context for each word
- Same process in reverse: right context for each word

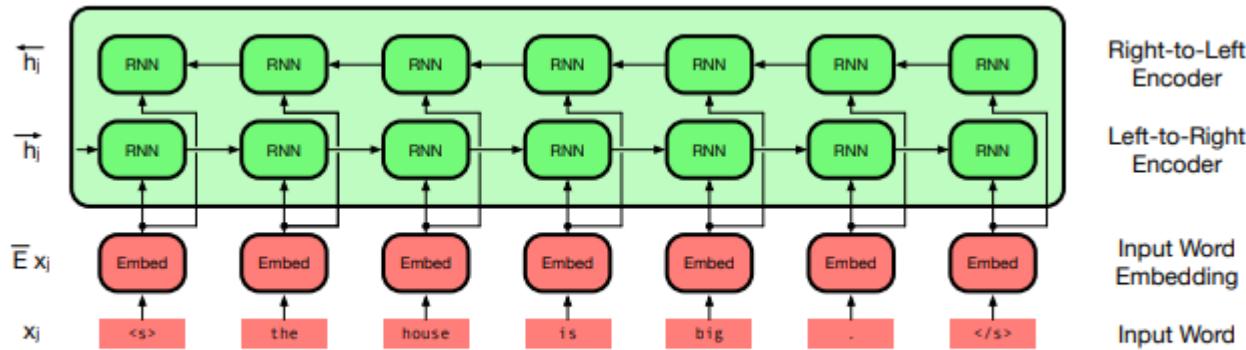


Input Encoder-Bidirectional RNN



- Input encoder: concatenate bidirectional RNN states
- Each word representation includes full left and right sentence context

Encoder Math



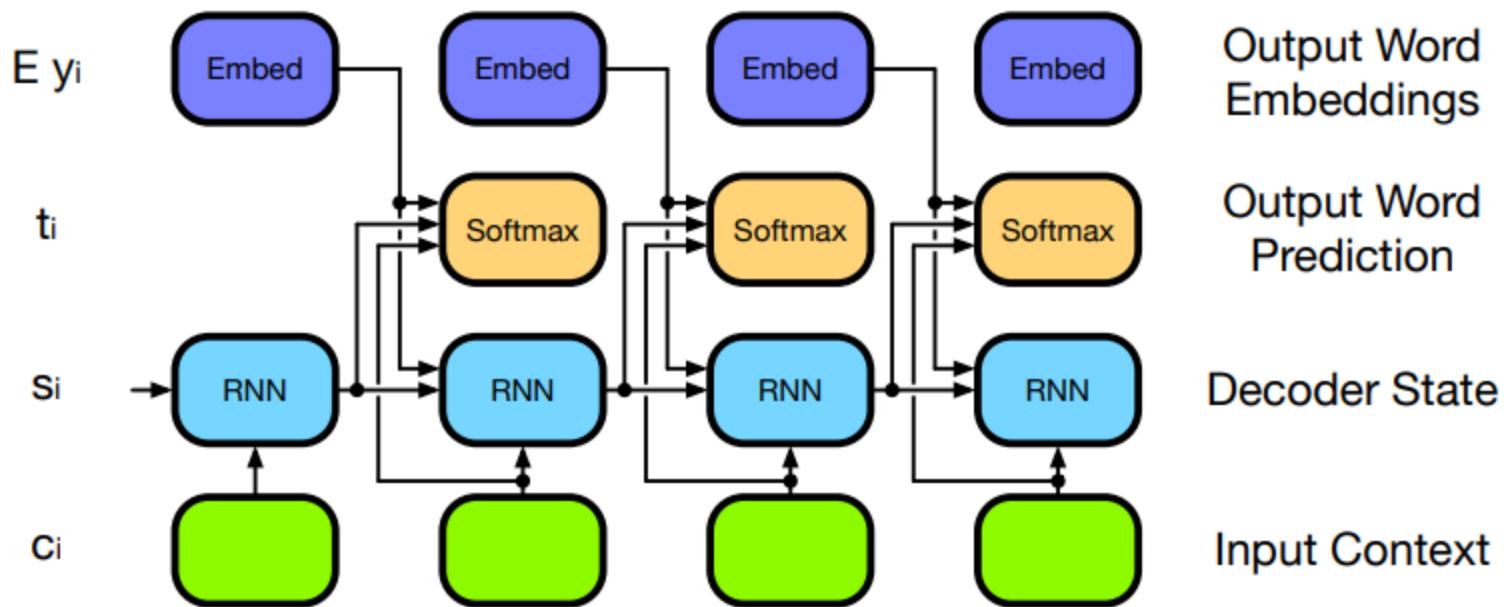
- Input is sequence of words x_j , mapped into embedding space $\bar{E} x_j$
- Bidirectional recurrent neural networks

$$\begin{aligned}\overleftarrow{h}_j &= f(\overleftarrow{h}_{j+1}, \bar{E} x_j) \\ \overrightarrow{h}_j &= f(\overrightarrow{h}_{j-1}, \bar{E} x_j)\end{aligned}$$

- Various choices for the function $f()$: feed-forward layer, GRU, LSTM, ...

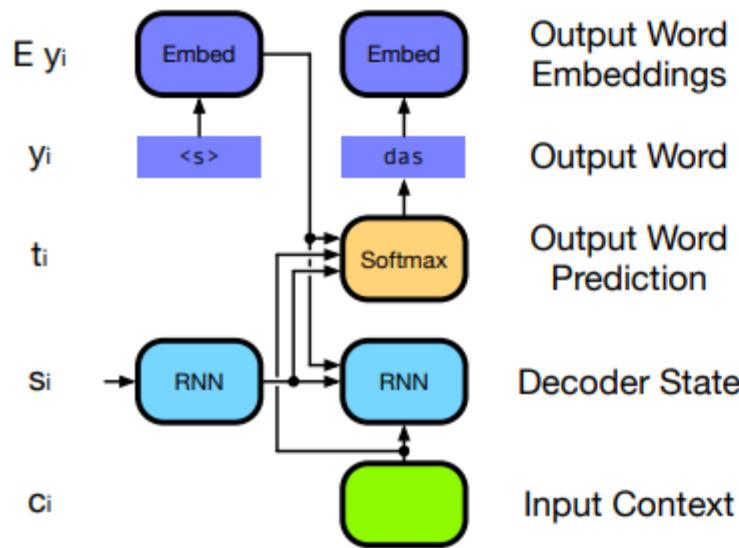
Decoder

- We want to have a recurrent neural network predicting output words



- We feed decisions on output words back into the decoder state
- Decoder state is also informed by the input context

Decoder



- Decoder is also recurrent neural network over sequence of hidden states s_i

$$s_i = f(s_{i-1}, E y_{i-1}, c_i)$$

- Again, various choices for the function $f()$: feed-forward layer, GRU, LSTM, ...
- Output word y_i is selected by computing a vector t_i (same size as vocabulary)

$$t_i = W(U s_{i-1} + V E y_{i-1} + C c_i)$$

then finding the highest value in vector t_i

- If we normalize t_i , we can view it as a probability distribution over words
- $E y_i$ is the embedding of the output word y_i

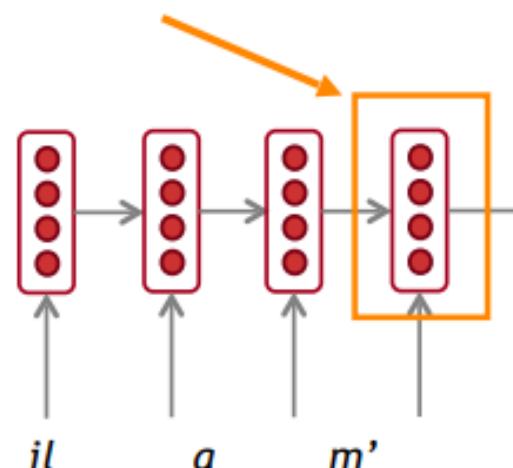
Issues with Encoder-Decoder

- A potential issue with this encoder-decoder approach is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector.
- This may make it difficult for the neural network to cope with long sentences.
- The performance of a basic encoder-decoder deteriorates rapidly as the length of an input sentence increases

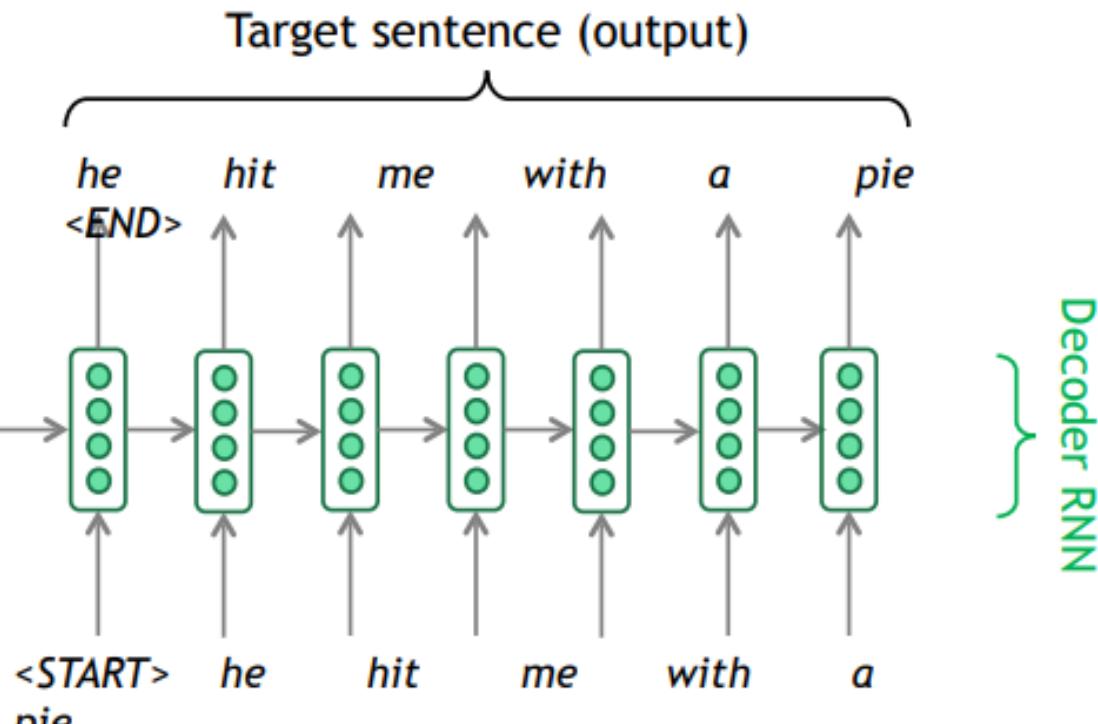
Bottleneck

Encoding of the source sentence.
This needs to capture *all information* about the source sentence.
Information bottleneck!

Encoder RNN



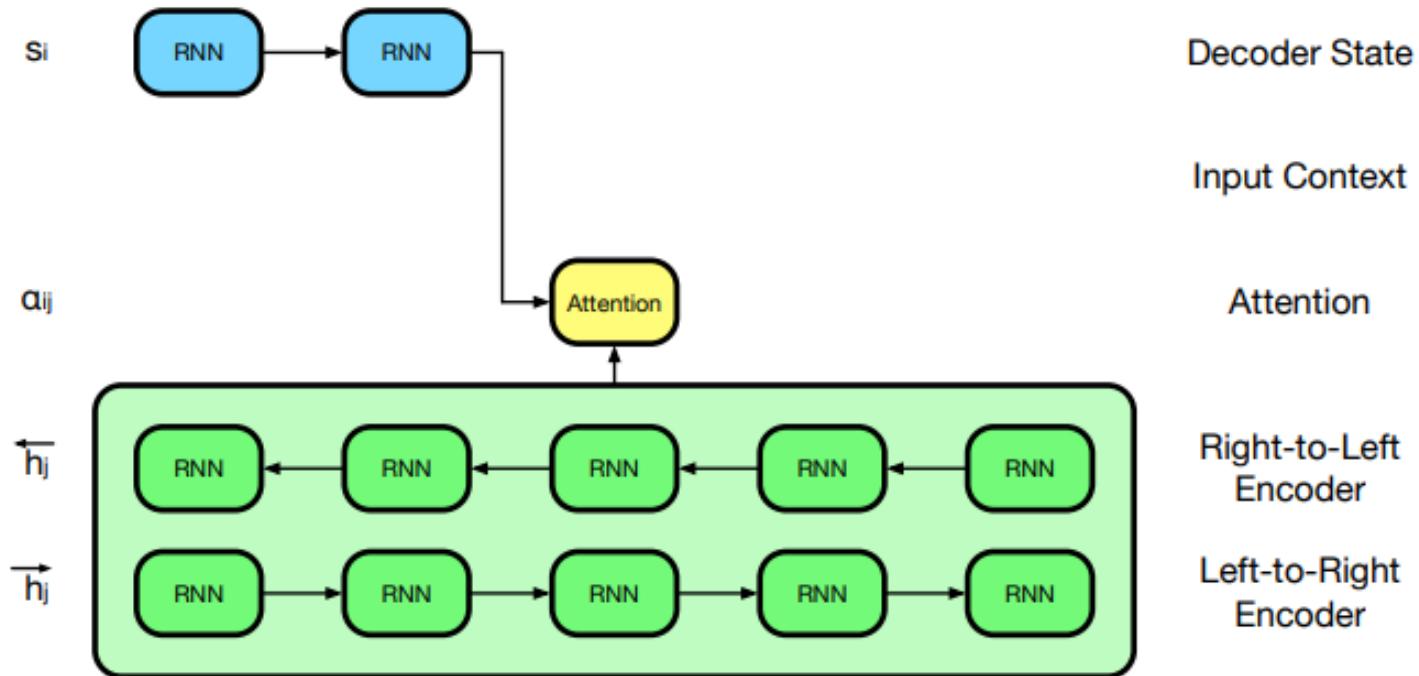
Source sentence (input)



Attention

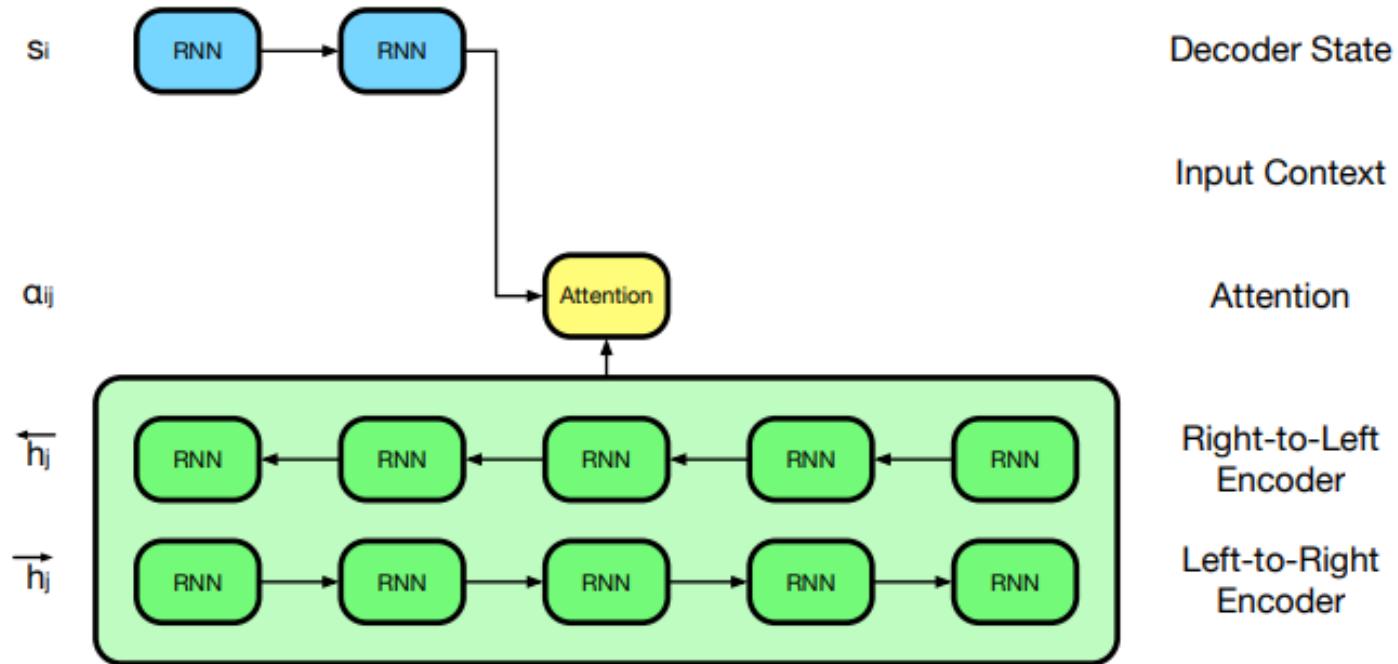
- Attention provides a solution to the bottleneck problem.
- Core idea: on each step of the decoder, use direct connection to the encoder to focus on a particular part of the source sequence
- First we will show via diagram (no equations), then we will show with equations

Attention



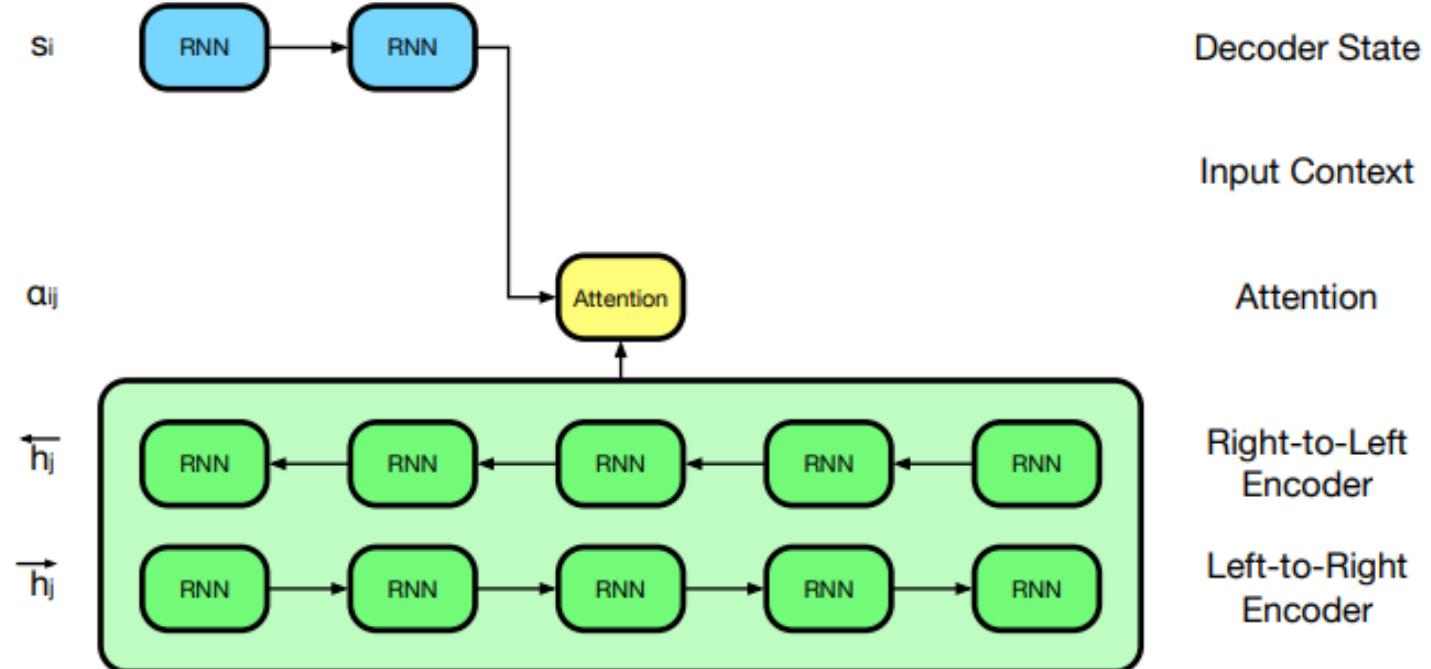
- Given what we have generated so far (decoder hidden state)
- ... which words in the input should we pay attention to (encoder states)?

Attention



- Given:
 - the previous hidden state of the decoder s_{i-1}
 - the representation of input words $h_j = (\overleftarrow{h_j}, \overrightarrow{h_j})$
- Predict an alignment probability $a(s_{i-1}, h_j)$ to each input word j (modeled with a feed-forward neural network layer)

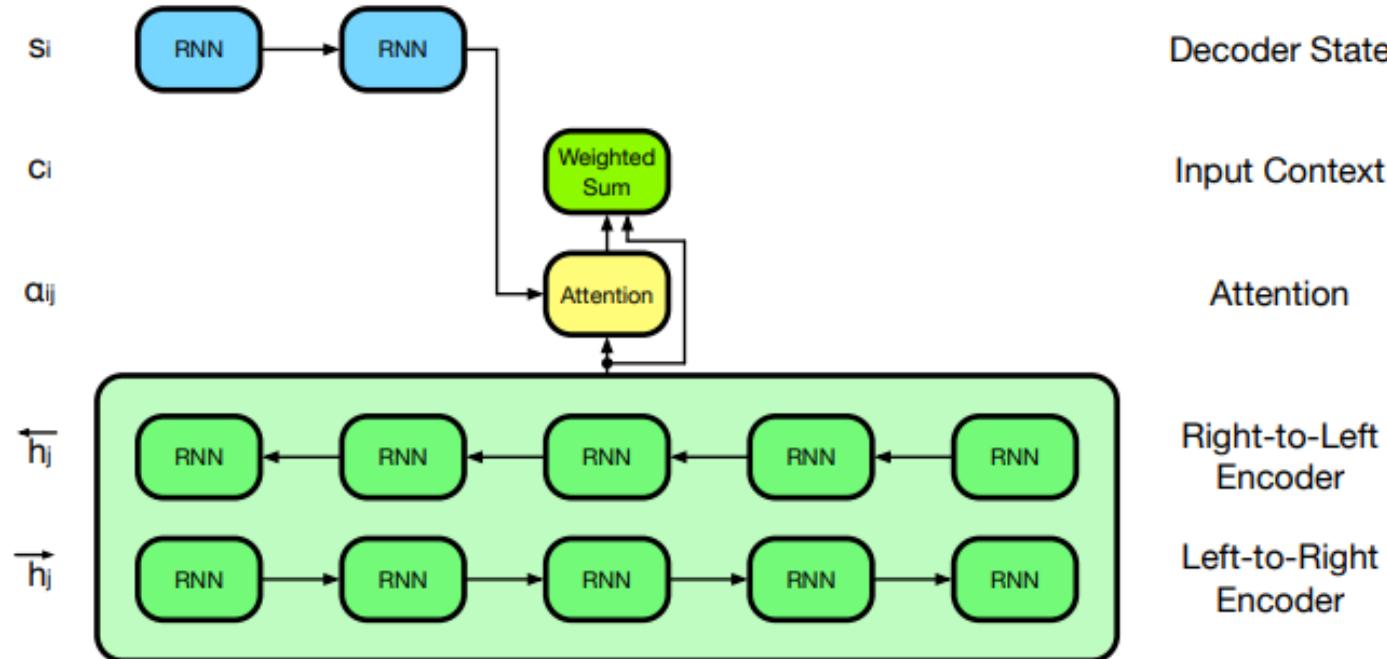
Attention



- Normalize attention (softmax)

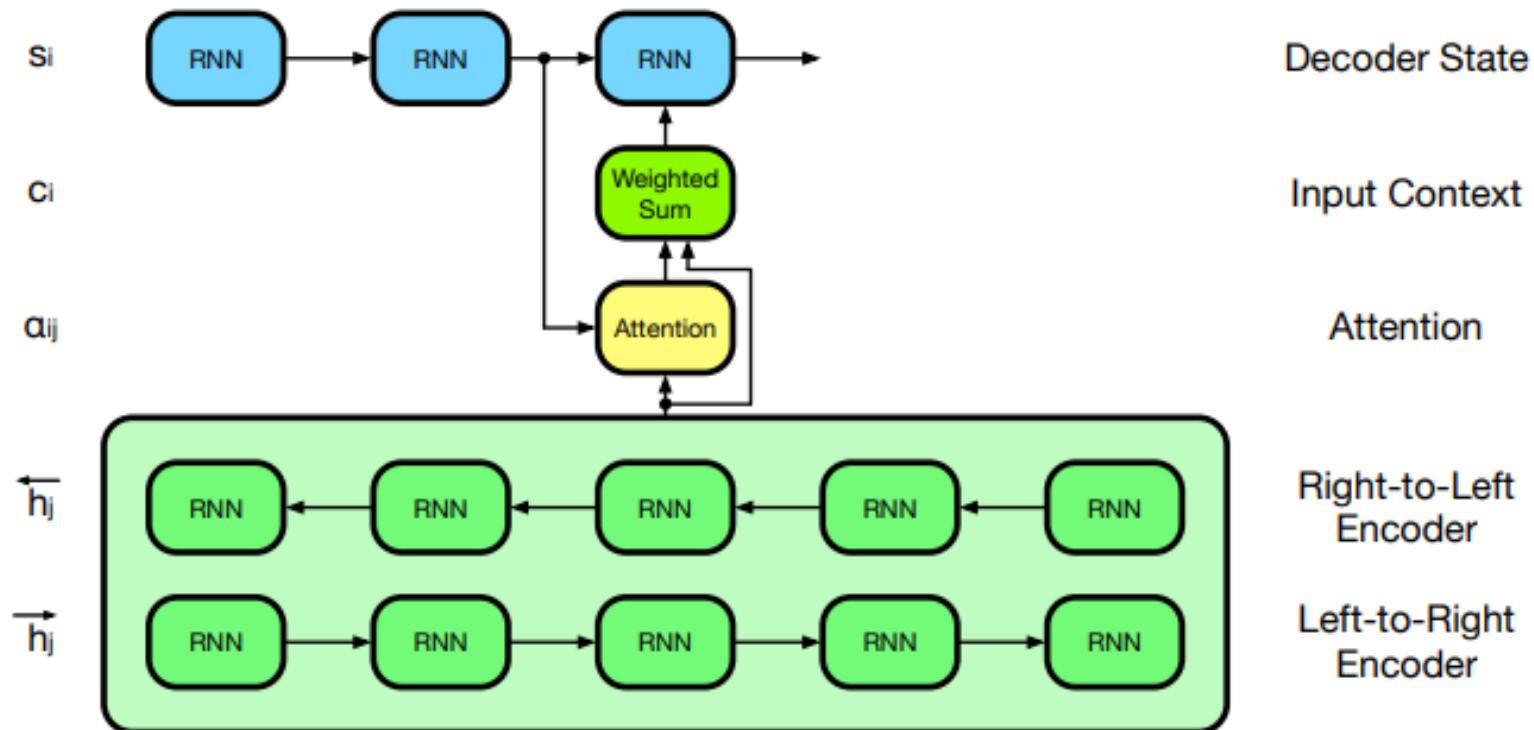
$$\alpha_{ij} = \frac{\exp(a(s_{i-1}, h_j))}{\sum_k \exp(a(s_{i-1}, h_k))}$$

Attention



- Relevant input context: weigh input words according to attention: $c_i = \sum_j \alpha_{ij} h_j$

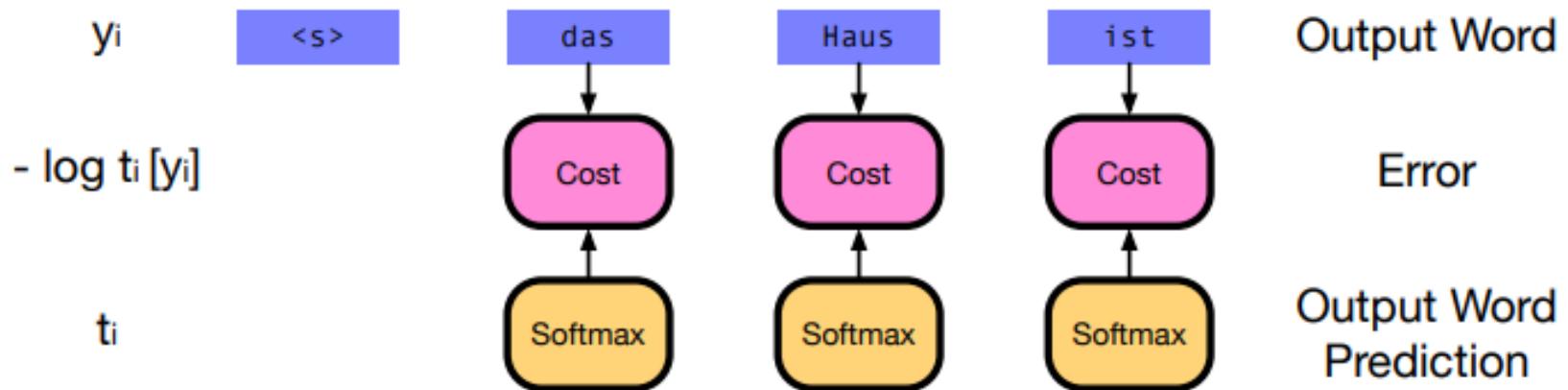
Attention



Use context to predict next hidden state and output word

Training Phase

Comparing predicted word with actual

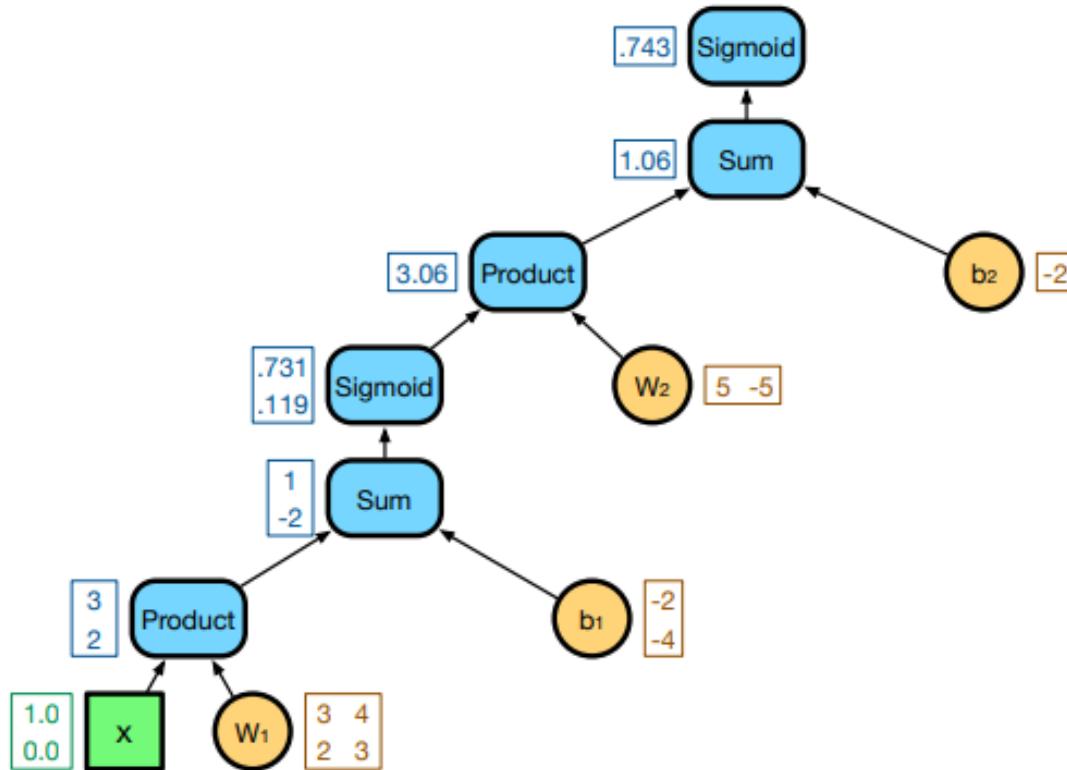


Current model gives some probability $t_i[y_i]$ to correct word y_i

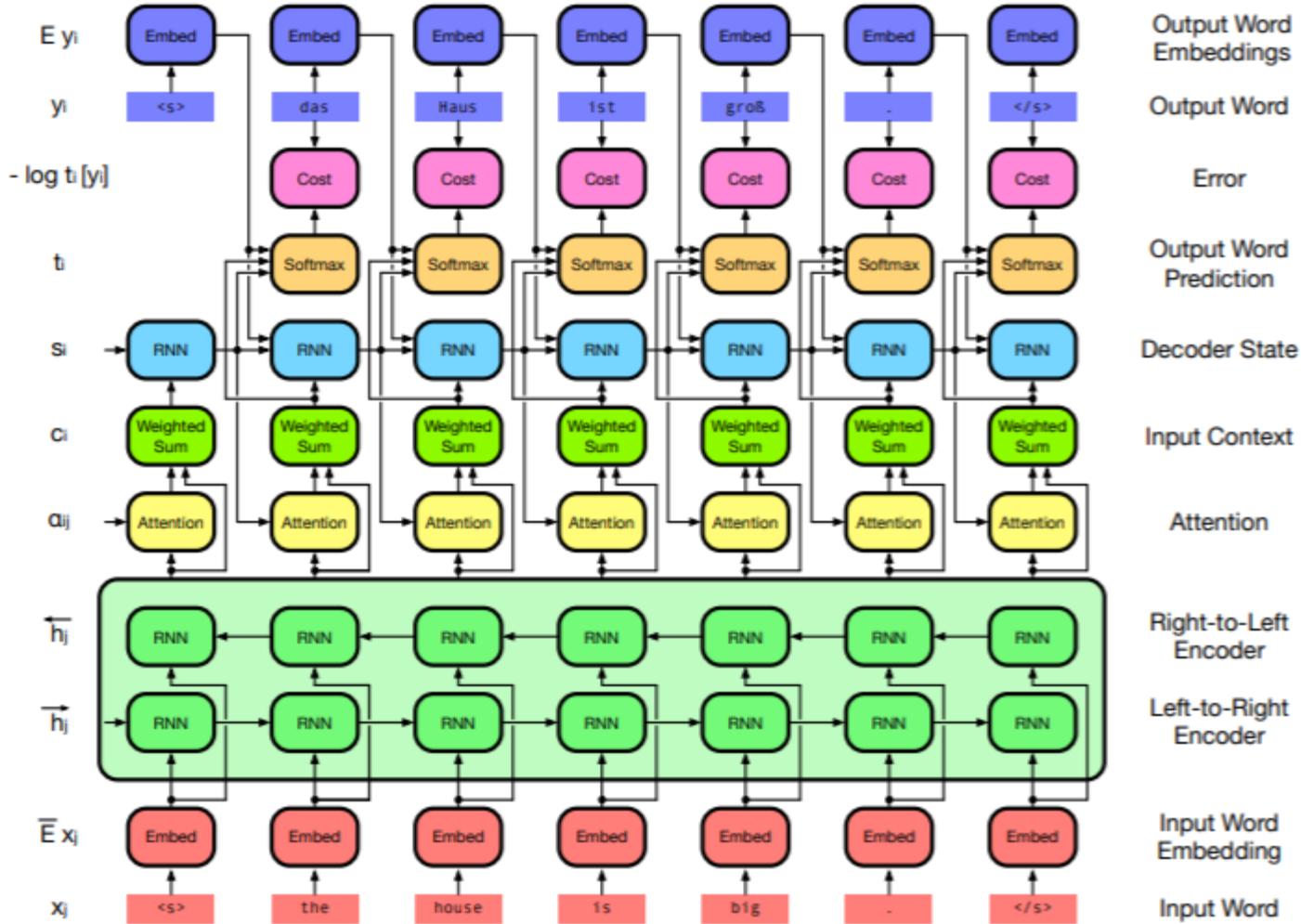
We turn this into an error by computing cross-entropy: $-\log t_i[y_i]$

Computation graph

- Math behind neural machine translation defines a computation graph
- Forward and backward computation to compute gradients for model training



Encoder-Decoder Model



Batching

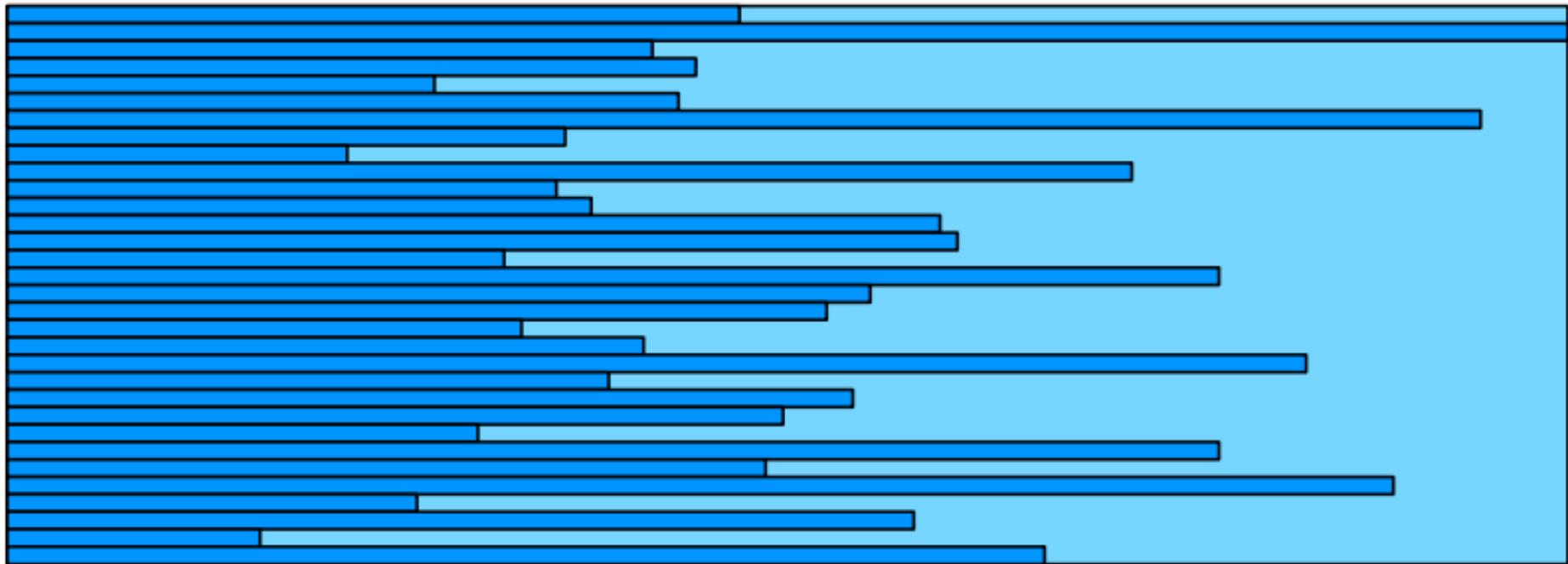
- Already large degree of parallelism
 - most computations on vectors, matrices
 - efficient implementations for CPU and GPU •

Further parallelism by batching

- - processing several sentence pairs at once
- Typical batch sizes 50-100 sentence pairs

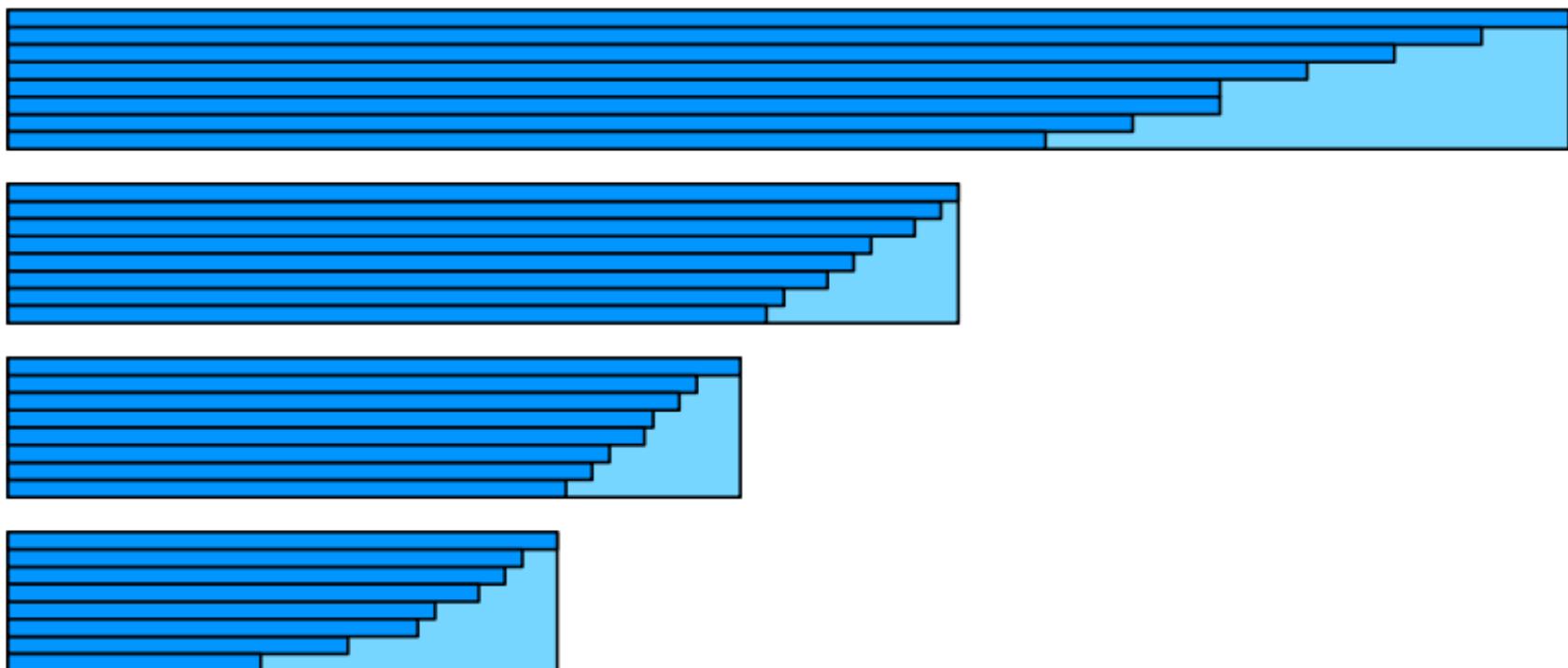
Batching

- Sentences differ in length
- A lot of wasted computations



Batching

- Sort sentences by length, break up into mini-batches
- Example: Maxi-batch 1600 sentence pairs, mini-batch 80 sentence pairs

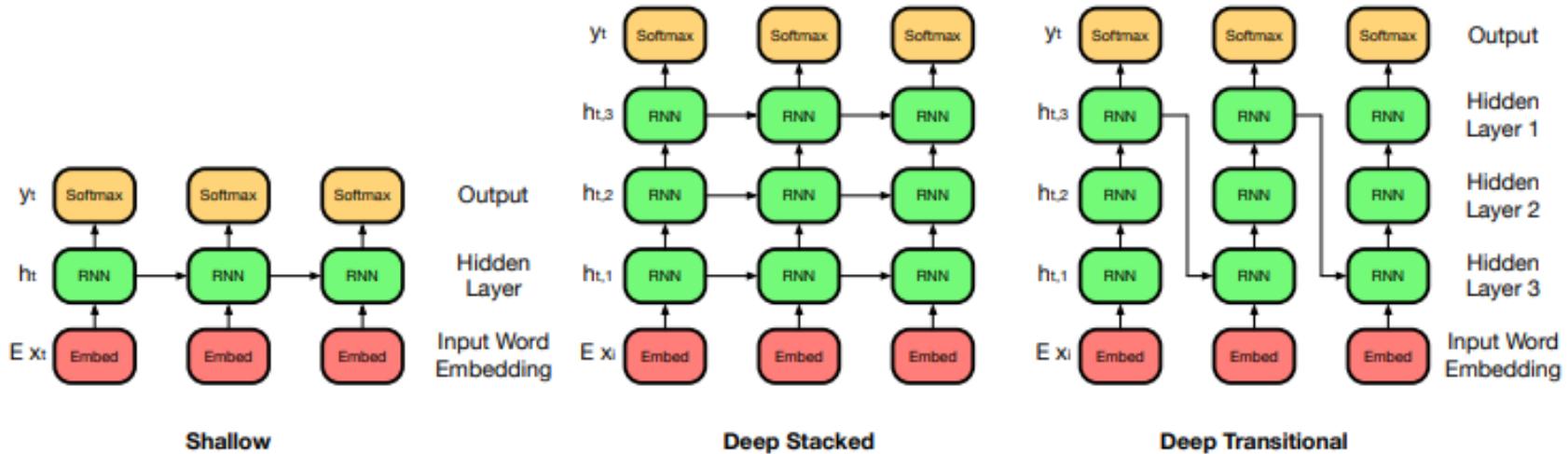


Organization of training

- Shuffle corpus
- Break into maxi-batches
- Break up each maxi-batch into mini-batches
- Process mini-batch, update parameters
- Once done, repeat
- Typically 5-15 epochs needed (passes through entire training corpus)

Deep network models

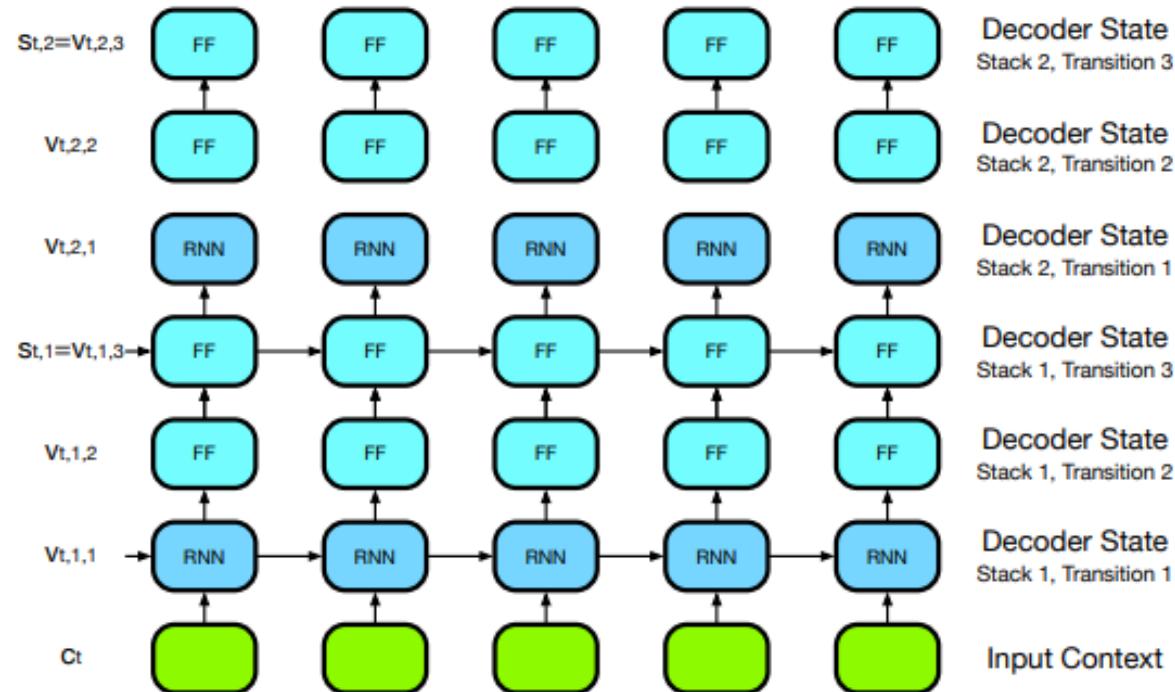
- Encoder and decoder are recurrent neural networks
- We can add additional layers for each step
- Recall shallow and deep language models



- Adding residual connections (short-cuts through deep layers) help

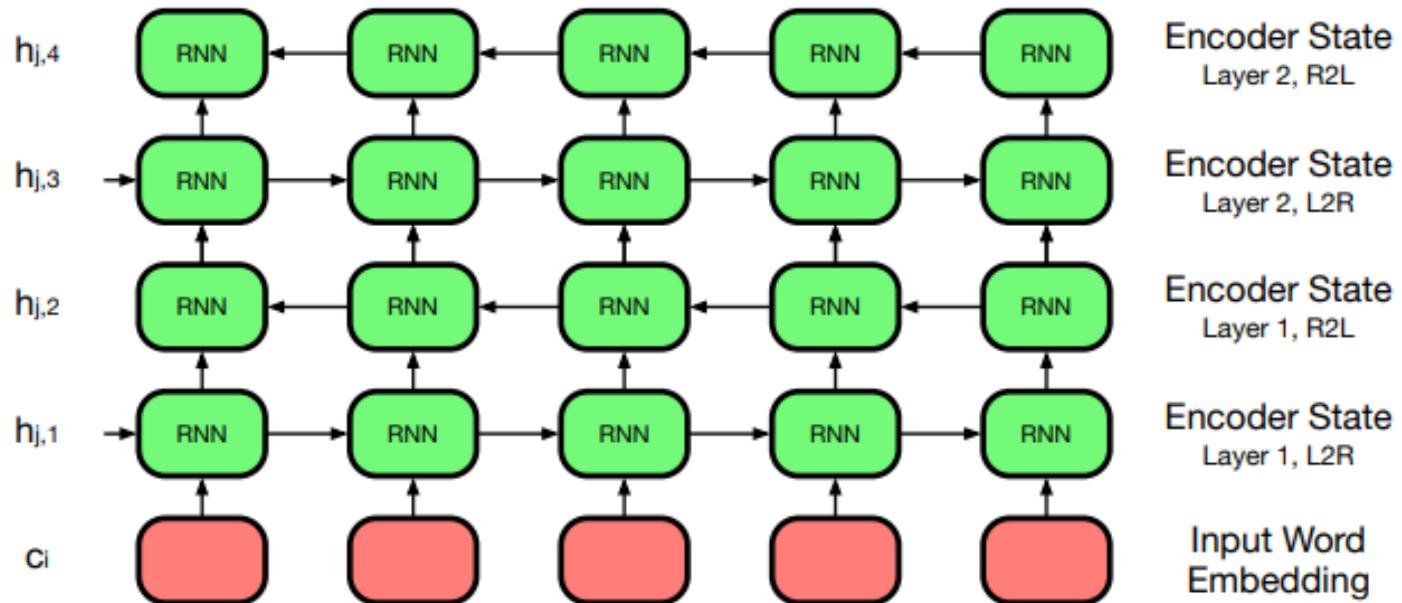
Deep Decoder

- Two ways of adding layers
 - deep transitions: several layers on path to output
 - deeply stacking recurrent neural networks
- Why not both?



Deep Encoder

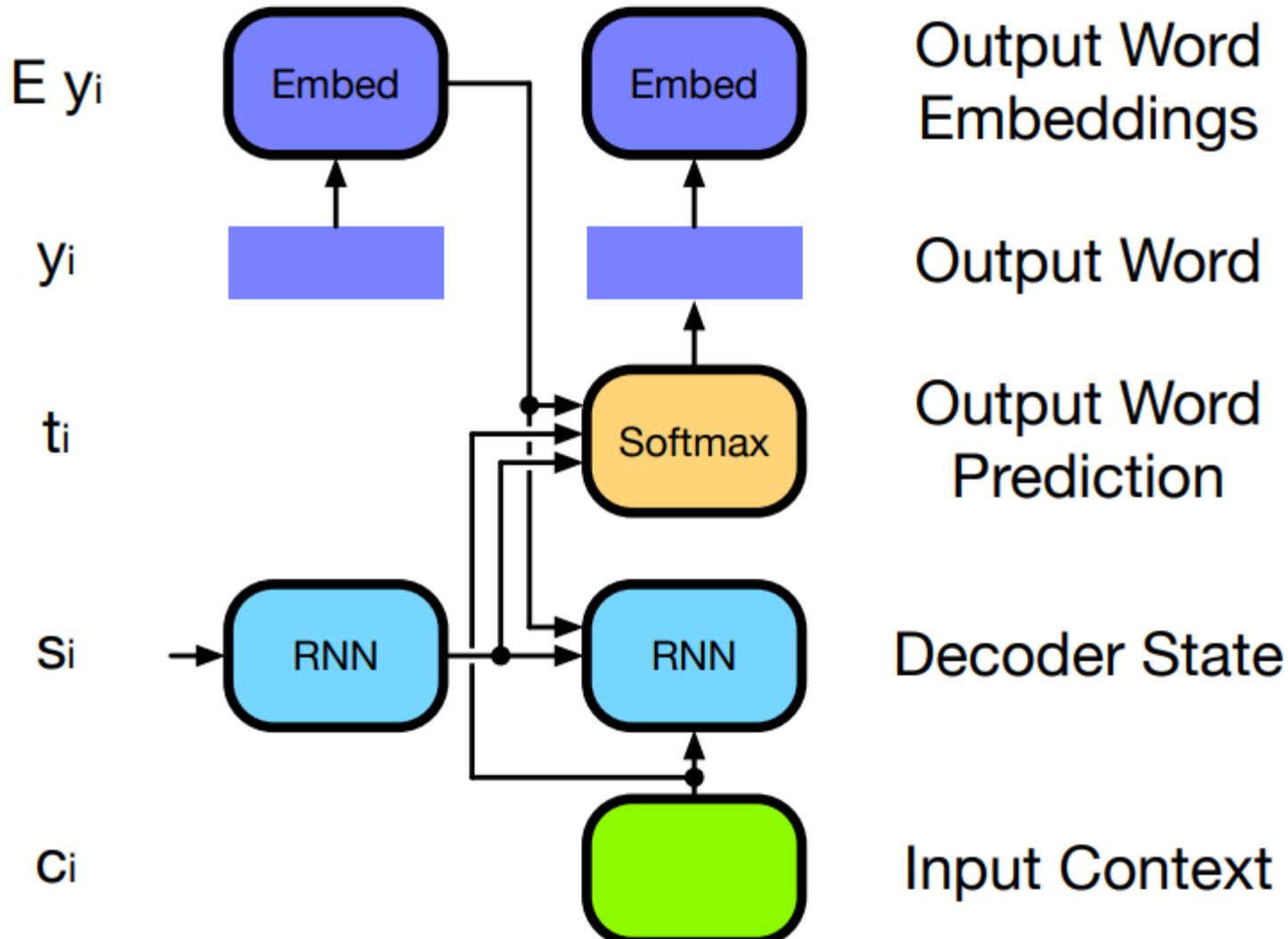
- Previously proposed encoder already has 2 layers
 - left-to-right recurrent network, to encode left context
 - right-to-left recurrent network, to encode right context
- ⇒ Third way of adding layers



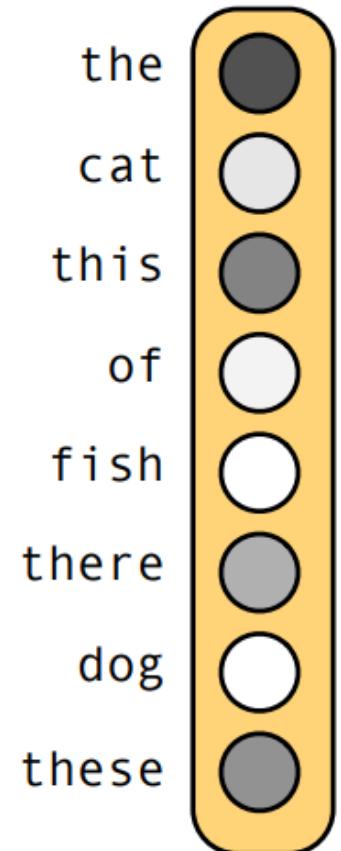
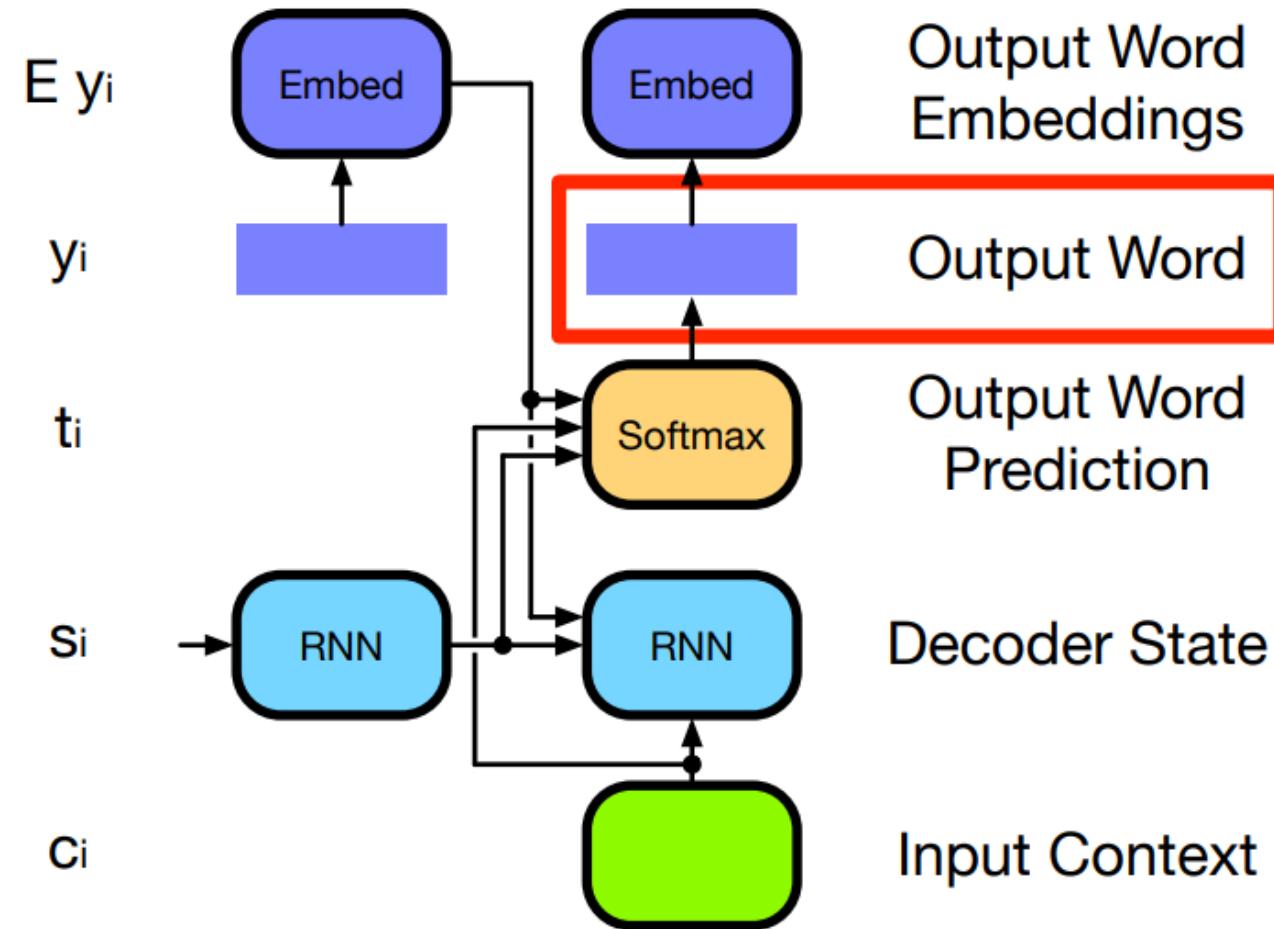
Inference

- Given a trained model
 - ... we now want to translate test sentences
- We only need execute the "forward" step in the computation graph

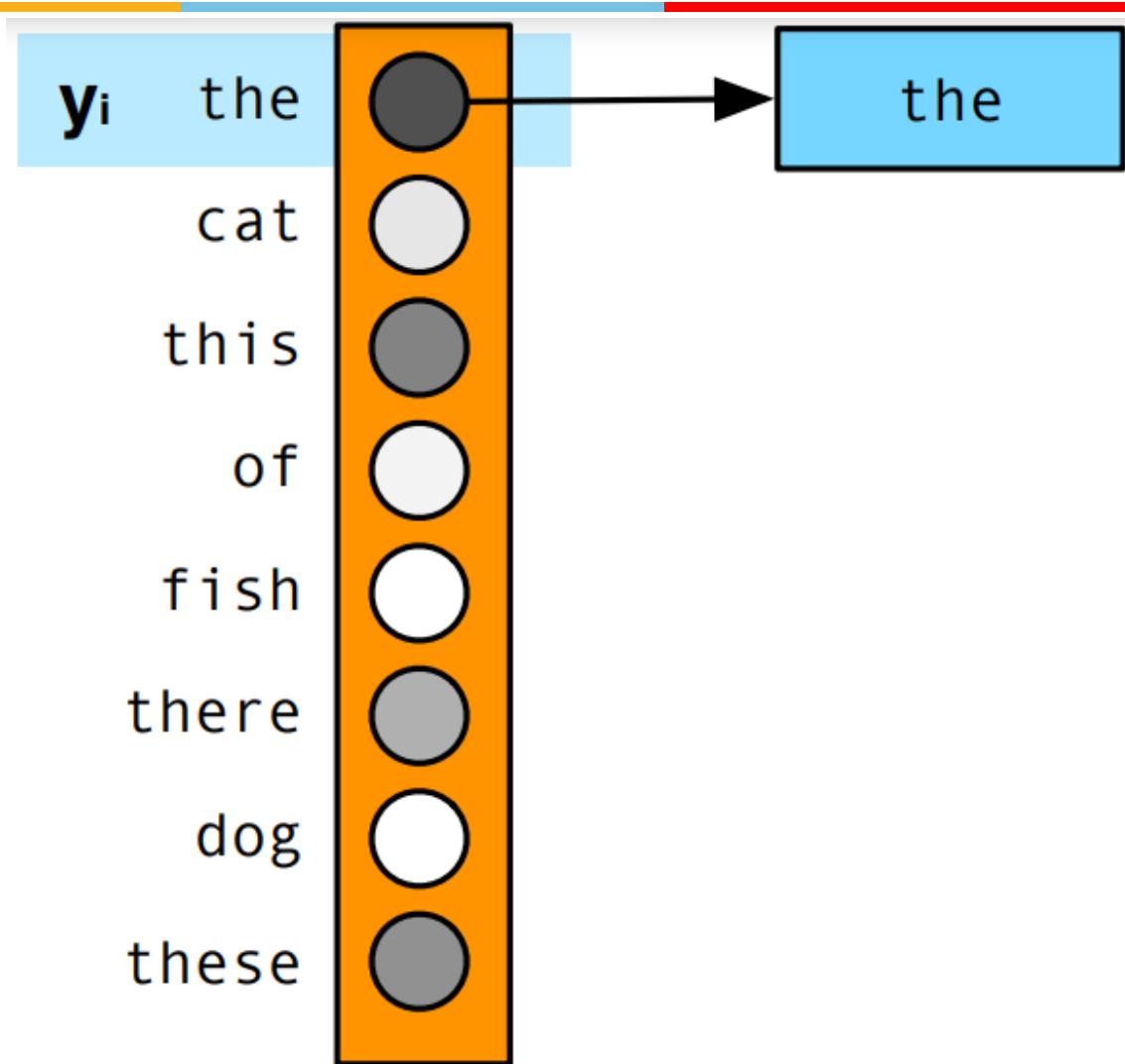
Word Prediction



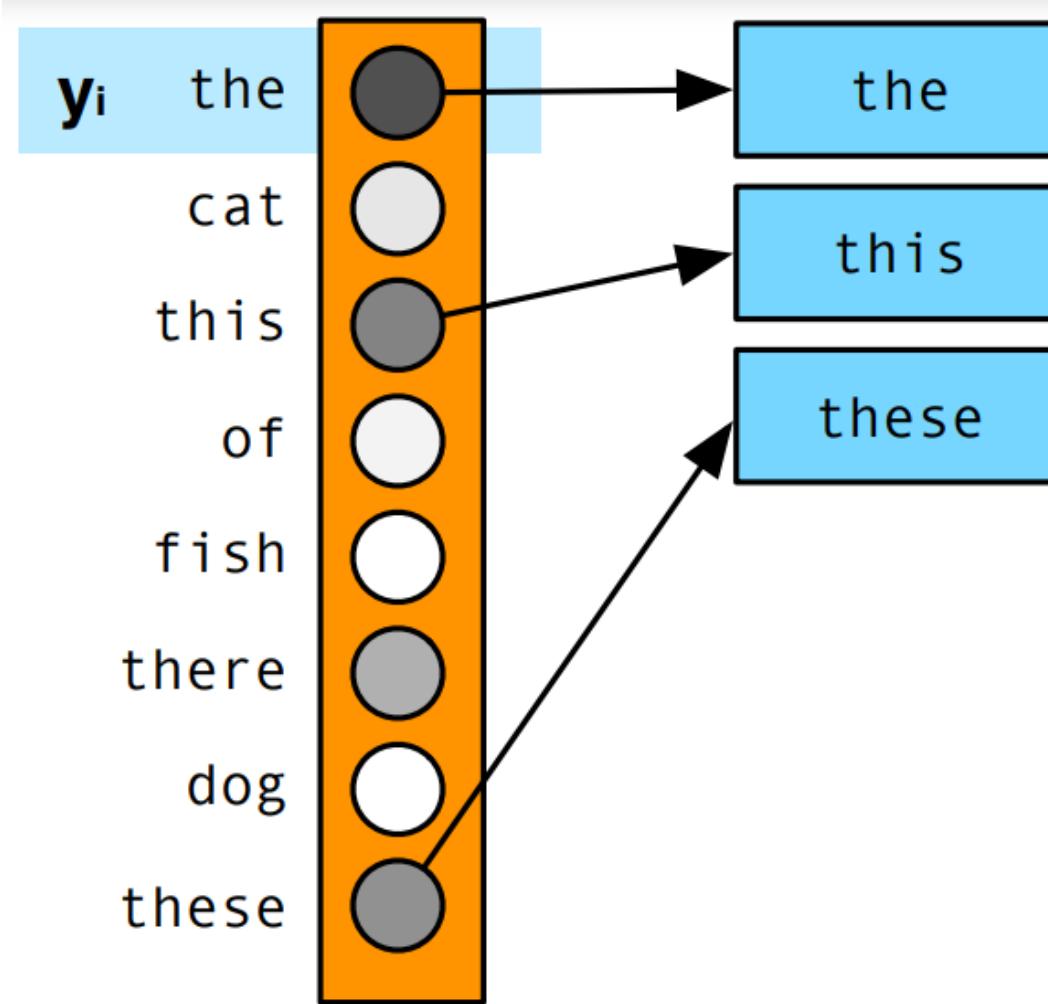
Selected word



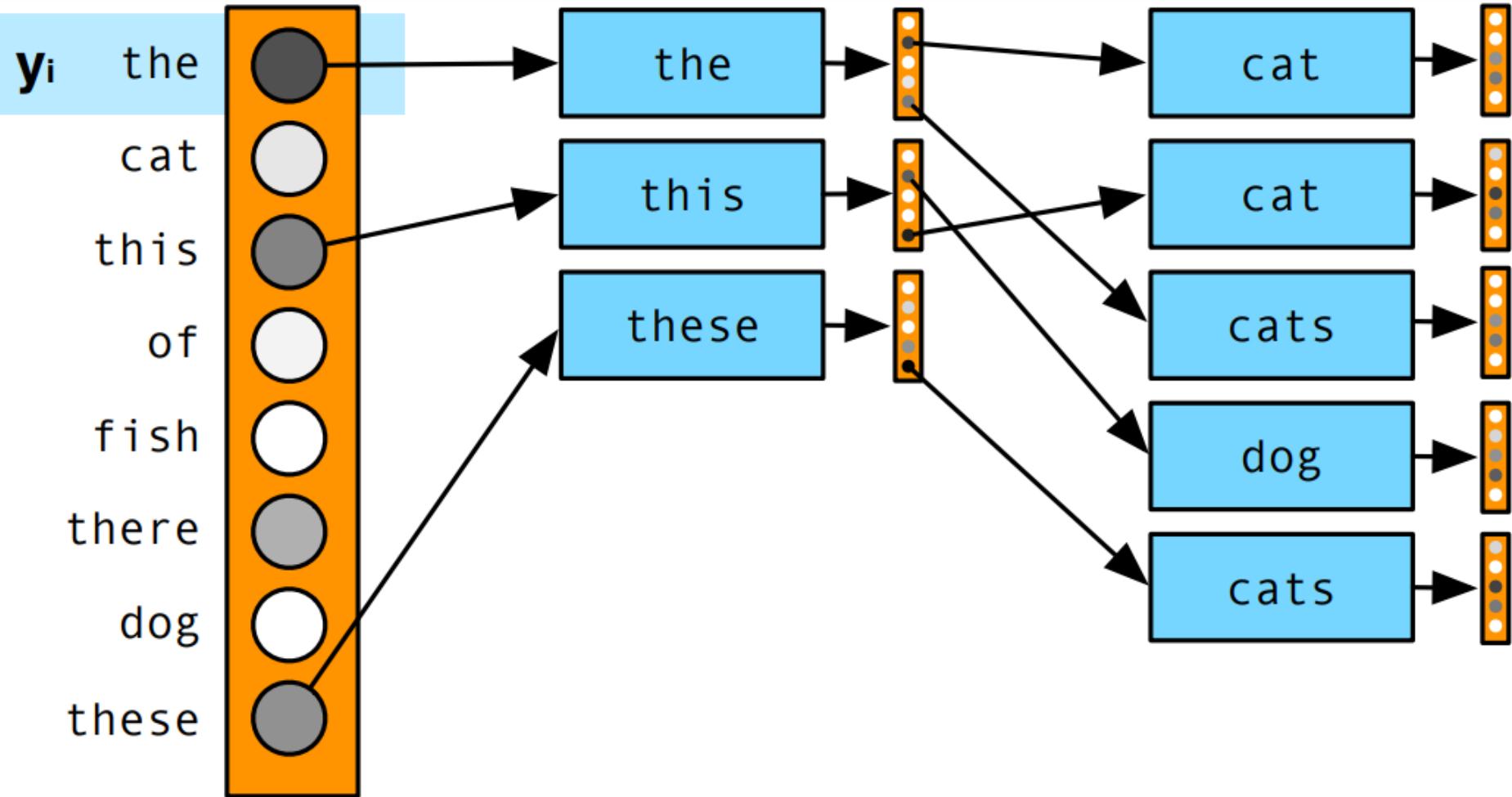
Select the best word



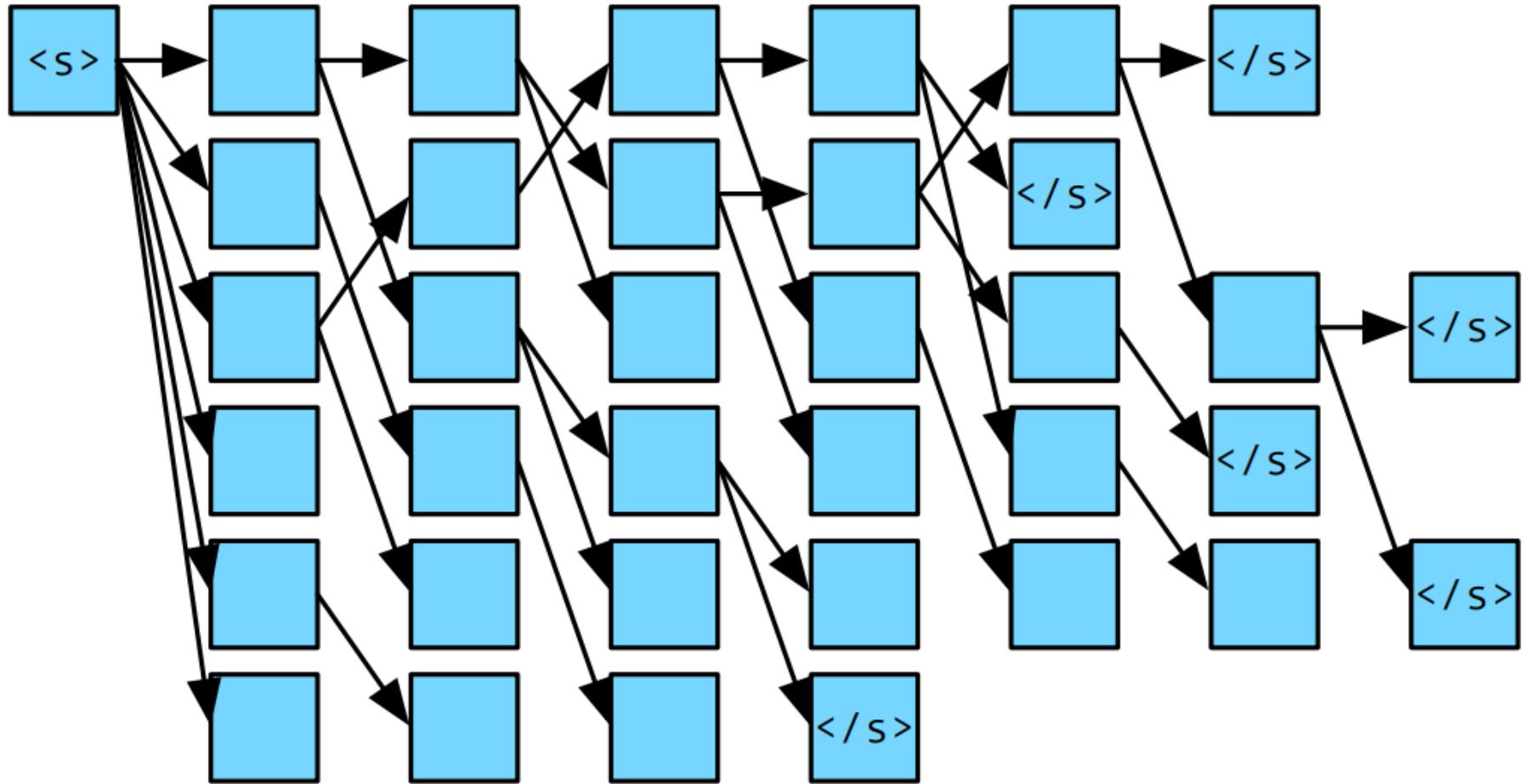
Select second and third best



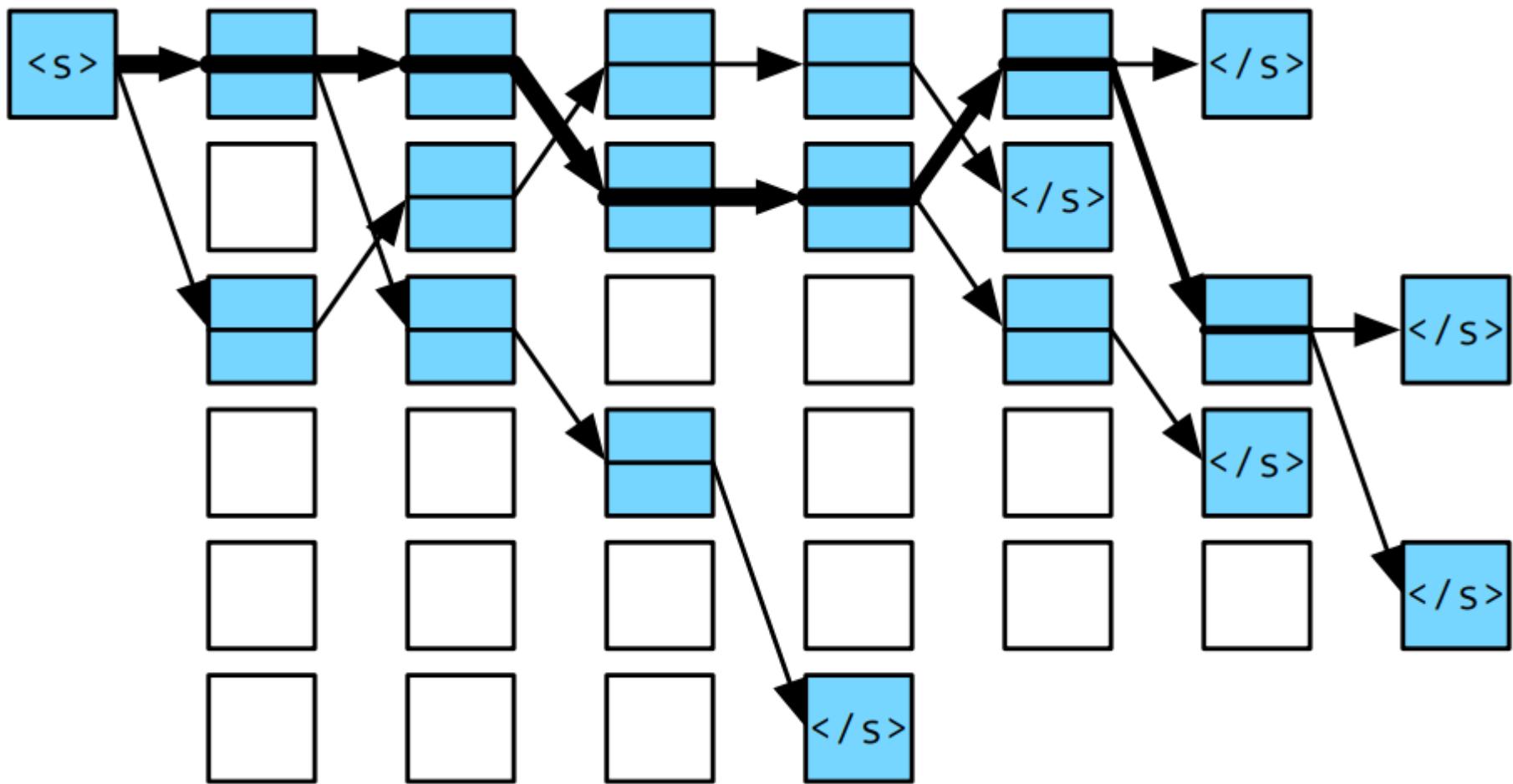
Select best continuation



Beam Search



Best path

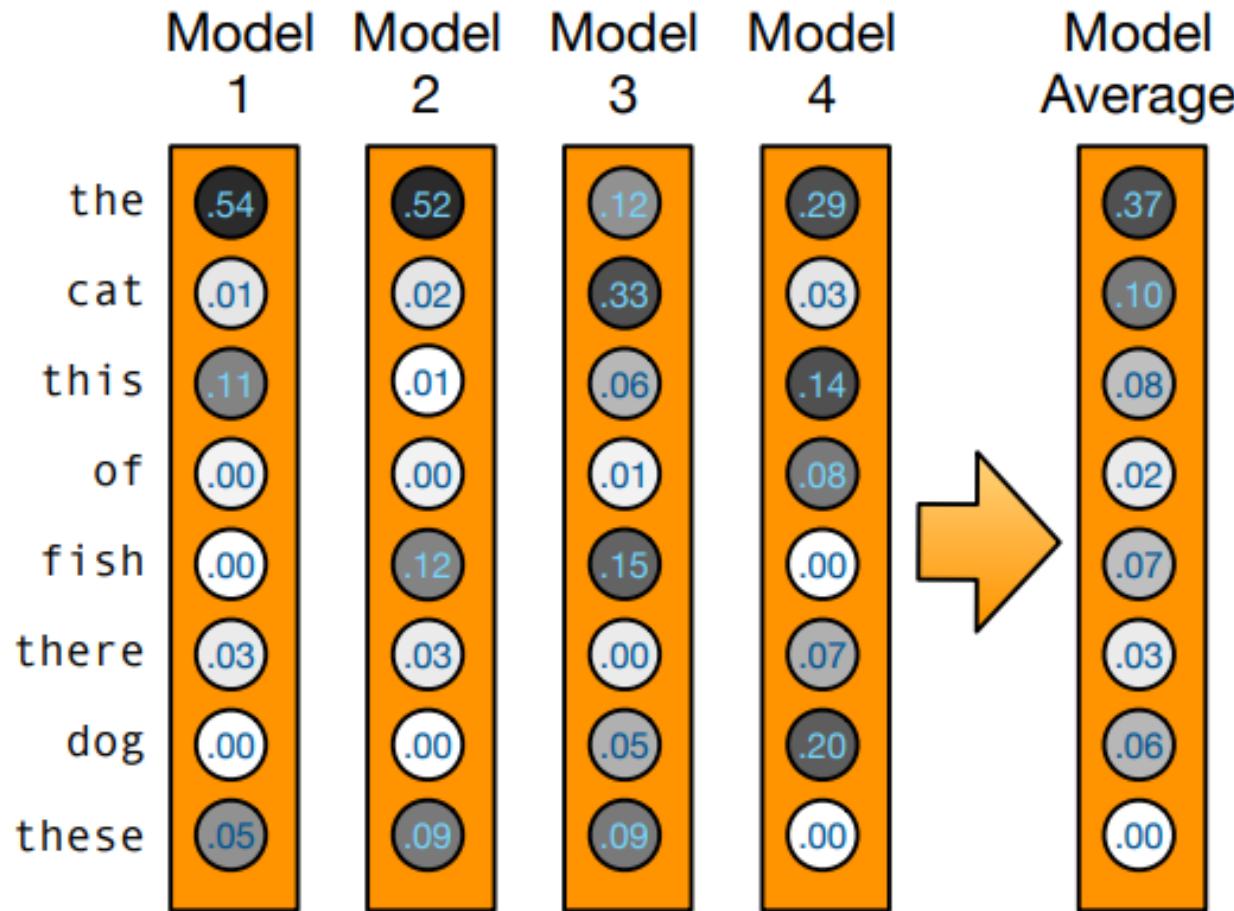


Output word predictions

Input Sentence: ich glaube aber auch , er ist clever genug um seine Aussagen vage genug zu halten , so dass sie auf verschiedene Art und Weise interpretiert werden können .

Best	Alternatives
but (42.1%)	however (25.3%), I (20.4%), yet (1.9%), and (0.8%), nor (0.8%), ...
I (80.4%)	also (6.0%), , (4.7%), it (1.2%), in (0.7%), nor (0.5%), he (0.4%), ...
also (85.2%)	think (4.2%), do (3.1%), believe (2.9%), , (0.8%), too (0.5%), ...
believe (68.4%)	think (28.6%), feel (1.6%), do (0.8%), ...
he (90.4%)	that (6.7%), it (2.2%), him (0.2%), ...
is (74.7%)	's (24.4%), has (0.3%), was (0.1%), ...
clever (99.1%)	smart (0.6%), ...
enough (99.9%)	about (1.2%), for (1.1%), in (1.0%), of (0.3%), around (0.1%), ...
to (95.5%)	maintain (4.5%), hold (4.4%), be (4.2%), have (1.1%), make (1.0%), ...
keep (69.8%)	its (2.1%), statements (1.5%), what (1.0%), out (0.6%), the (0.6%), ...
his (86.2%)	testimony (1.5%), messages (0.7%), comments (0.6%), ...
statements (91.9%)	v@@@ (1.2%), in (0.6%), ambiguous (0.3%), ...
vague (96.2%)	and (0.2%), ...
enough (98.9%)	, (44.3%), to (1.2%), in (0.6%), and (0.5%), just (0.2%), that (0.2%), ...
so (51.1%)	that (35.3%), it (2.5%), can (1.6%), you (0.8%), we (0.4%), to (0.3%), ...
they (55.2%)	may (2.7%), could (1.6%), are (0.8%), will (0.6%), might (0.5%), ...
can (93.2%)	have (0.3%), interpret (0.2%), get (0.2%), ...
be (98.4%)	interpre@@ (0.1%), constru@@ (0.1%), ...
interpreted (99.1%)	on (0.9%), differently (0.5%), as (0.3%), to (0.2%), for (0.2%), by (0.1%), ...
in (96.5%)	a (25.2%), various (22.7%), several (3.6%), ways (2.4%), some (1.7%), ...
different (41.5%)	way (0.2%), manner (0.2%), ...
ways (99.3%)	</s> (0.2%), , (0.1%), ...
.	
</s>	(100.0%)

Combine predictions



Ensembling

- Surprisingly reliable method in machine learning
- Long history, many variants: bagging, ensemble, model averaging, system combination, ...
- Works because errors are random, but correct decisions unique

Right to left and left to right generation

- Neural machine translation generates words right to left (L2R)

the → cat → is → in → the → bag → .

- But it could also generate them right to left (R2L)

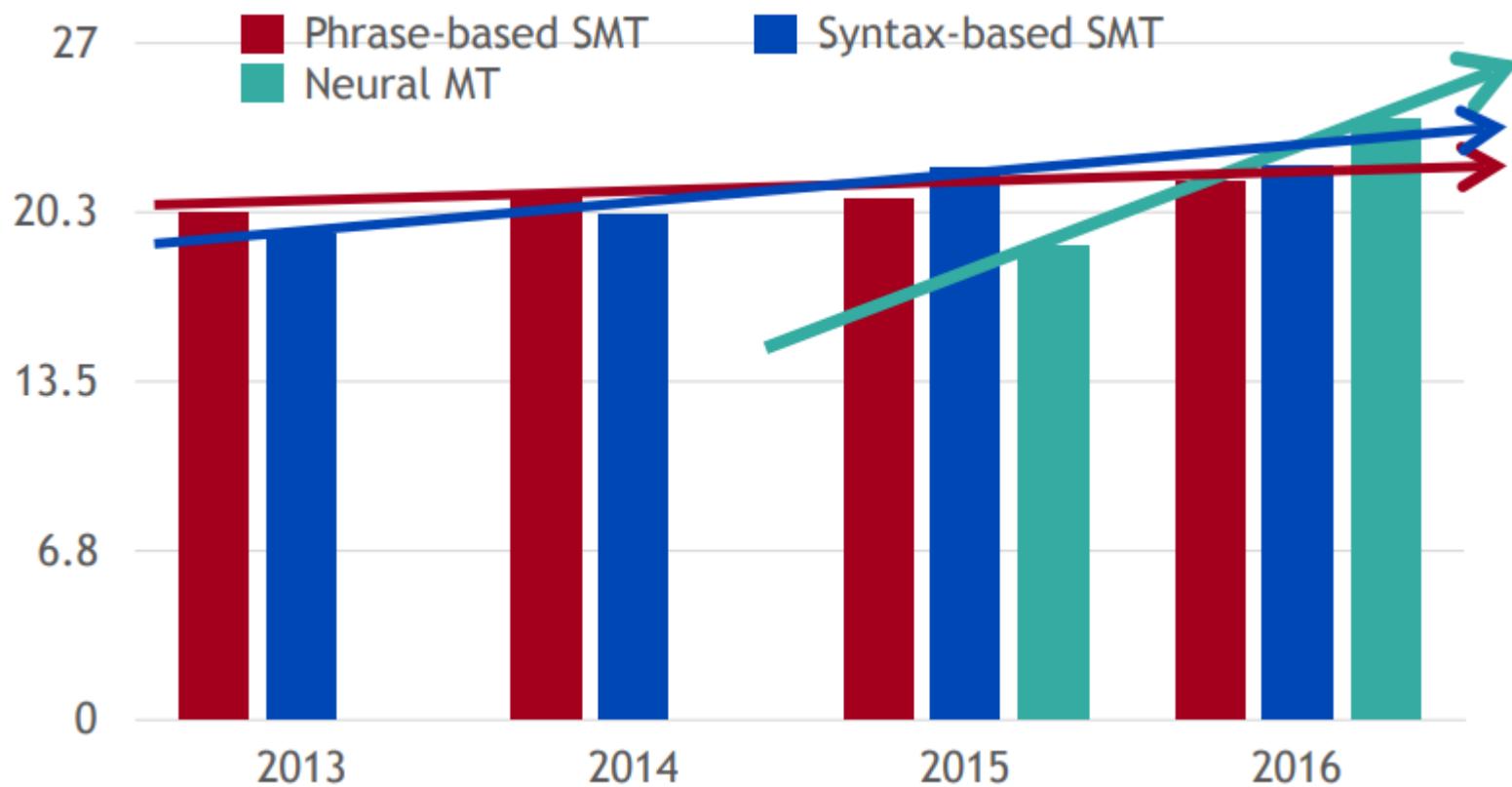
the ← cat ← is ← in ← the ← bag ← .

Obligatory notice: Some languages (Arabic, Hebrew, ...) have writing systems that are right-to-left, so the use of "right-to-left" is not precise here.

Reranking

- Train both L2R and R2L model
- Score sentences with both \Rightarrow use both left and right context during translation
- Only possible once full sentence produced \rightarrow re-ranking
 1. generate n-best list with L2R model
 2. score candidates in n-best list with R2L model
 3. chose translation with best average score

MT Progress over time



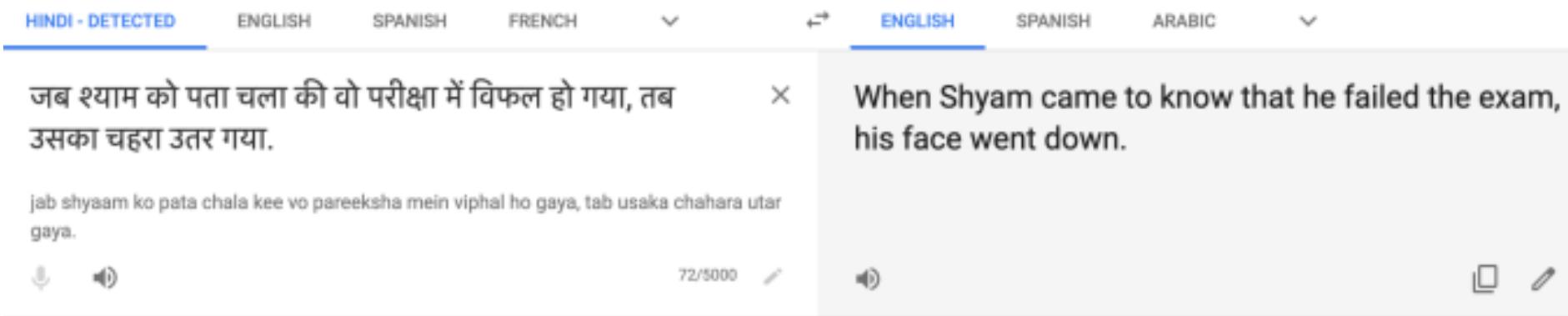
Is Machine Translation solved?

Nope!

- Many difficulties remain:
- Out-of-vocabulary words
- Domain mismatch between train and test data
- Maintaining context over longer text
- Low-resource language pairs

Is Machine Translation solved?

- Using common sense is still hard
- Idioms are difficult to translate



The screenshot shows a machine translation interface. On the left, the source text is in Hindi: "जब श्याम को पता चला की वो परीक्षा में विफल हो गया, तब उसका चहरा उतर गया." Below it is the raw text: "jab shyaam ko pata chala kee vo pareeksha mein viphal ho gaya, tab usaka chahara utar gaya." On the right, the target text is in English: "When Shyam came to know that he failed the exam, his face went down." There are also small icons for microphone, speaker, and edit.

HINDI - DETECTED	ENGLISH	SPANISH	FRENCH	▼	↔	ENGLISH	SPANISH	ARABIC	▼
जब श्याम को पता चला की वो परीक्षा में विफल हो गया, तब उसका चहरा उतर गया.	When Shyam came to know that he failed the exam, his face went down.								

Is Machine Translation solved?

NMT picks up biases in training data

Malay



English

dia bakerja sebagai
pengaturkara

he works as a
programmer

Malay



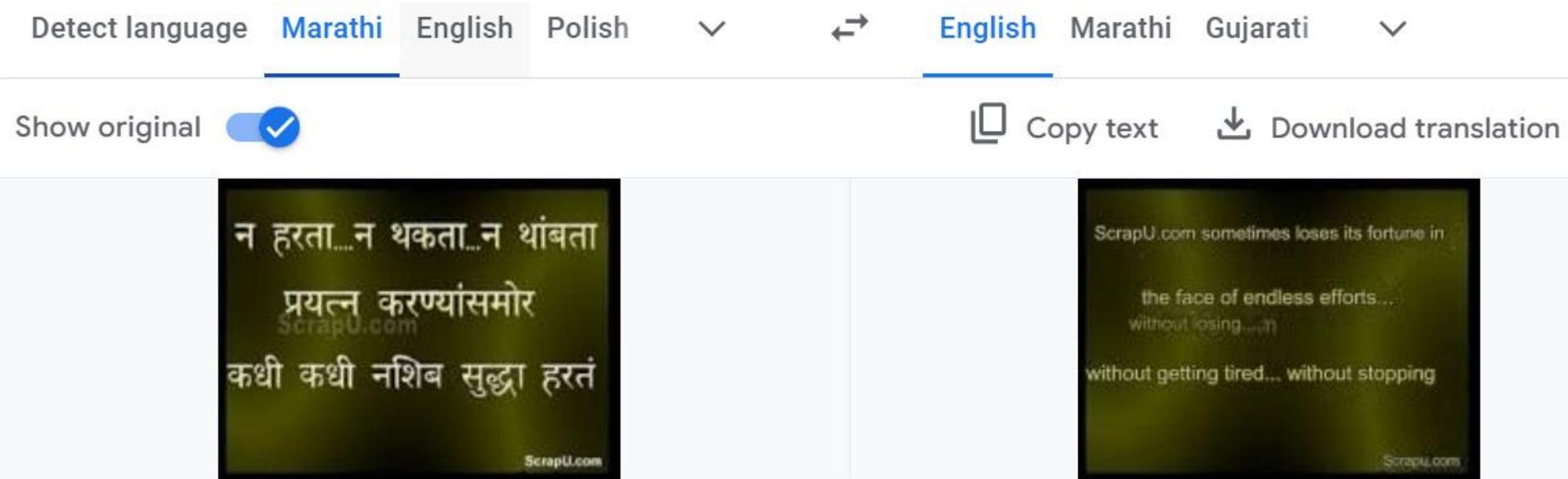
English

dia bakerja sebagai
jururawat

she works as a nurse

Is Machine Translation solved?

Uninterpretable systems do strange things



The screenshot shows a machine translation interface. At the top, there are language selection dropdowns for "Detect language" (set to Marathi), "Marathi" (selected), "English", "Polish", and "Gujarati". Below these are buttons for "Show original" (with a checked checkbox), "Copy text", and "Download translation". The left panel displays Marathi text from ScrapU.com:

न हरता...न थकता...न थांबता
प्रयत्न करण्यासमोर
कधी कधी नशिव सुद्धा हरतं

The right panel shows the English translation:

ScrapU.com sometimes loses its fortune in
the face of endless efforts...
without losing...
without getting tired... without stopping

NMT Advantages

Compared to SMT, NMT has many advantages:

- Better performance
 - More fluent
 - Better use of context
 - Better use of phrase similarities
 - A single neural network to be optimized end-to-end
 - No subcomponents to be individually optimized
 - Requires much less human engineering effort
 - No feature engineering
 - Same method for all language pairs
-

NMT Disadvantages

- NMT is less interpretable
- Hard to debug
- NMT is difficult to control
- For example, can't easily specify rules or guidelines for translation
- Safety concerns

Machine Learning Steps overview



A Machine learning pipeline typically consists of:

Importing Data: csv, xls, JSON

Exploratory Data Analysis

Data Pre-processing

Model Building

Model Evaluations



DEMO

Python Libraries

- **Pandas:** Data Manipulations
 - **NumPy:** Mathematical operations
 - **Scikit-learn:** Scikit-learn is one of the most popular ML libraries for classical ML algorithms
 - **Matplotlib and Seaborn:** Visualizations
 - **Tensorflow:** Deep Learning Library
-

Python program using Keras(Neural Network)

- Algorithm/pipeline
 - 1. Investigating the dataset for training
 - 2. Preprocessing of data
 - 2.1 converting text to integers
 - 2.2 Tokenization
 - 2.3 Padding
 - 3. Model selection
 - 4. Training the model

Case study

- Input : English sentence
- Output : Marathi sentence
- Approach used LSTM
- Simple MT is done with converting word in to integer
- Word embedding can be used to convert the word in to numvector

References

- Neural Machine Translation by Philip Koehn
- <https://www.youtube.com/watch?v=1uoOk2S6GUk>
- <https://www.youtube.com/watch?v=0DsWLXNIxeA&t=188s>
- Neural Machine Translation by Manning
- <https://www.youtube.com/watch?v=IxQtK2SjWWM>
- [Google's multilingual neural machine translation system: Enabling zero-shot translation](#)
[M Johnson](#), [M Schuster](#), [QV Le](#), M Krikun, [Y Wu](#), Z Chen, N Thorat, [F Viégas](#), [M Wattenberg](#)...
- Transactions of the Association for Computational Linguistics, 2017 • direct.mit.edu
- <https://www.youtube.com/watch?v=AlpXjFwVdIE>



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 11: Machine translation- Indic Languages

Date – 25th February 2023

These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philipp Koehn**, Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

Agenda

Indic Machine translation

- Introduction
- Recurrent Neural Translation Models
- Encoder Decoder Translation Models
- Neural translation model with attention
- Training Neural Models
- Deeper Models
- Demo

Is there gain in knowledge or loss of Knowledge?

- *From an estimated 10,000 world languages in 1900, about 6,700 language survived in 2000. Two percent of the world's languages are becoming extinct every year.*
 - *There is worldwide, unquantifiable erosion of cultural participation, knowledge and innovation.*
 - *With the loss of a language, we lose art and ideas, scientific information and technological innovation capacity.*
 - *World-level literacy is improving. More people can read than ever before, but fewer people create stories.*
 - *The share held by top four translated languages (English, Spanish, French and German) rose from 65 percent in 1980 to 81 percent in 1994.*
-
- ❖ ***Erosion of Language and Culture !!***

Is the technology to divide or to unite ?

- Latin Alphabet users , 39 % of the global population enjoy 84% of access to the Internet
- Hanzi-users in (CJK), 22% in global population enjoy 13% of Internet access
- Arbic script users, 9% of the population have 1.2 % of the Internet Access
- Bralmi-origin scripts users in South-east Asia and Indic scripts users occupy 22 % of the World population have just 0.3 % of Internet access.
- More than 80% content on Internet is in English.
- ICT penetration in India and other developing countries is lower.

Indic Machine Translation?

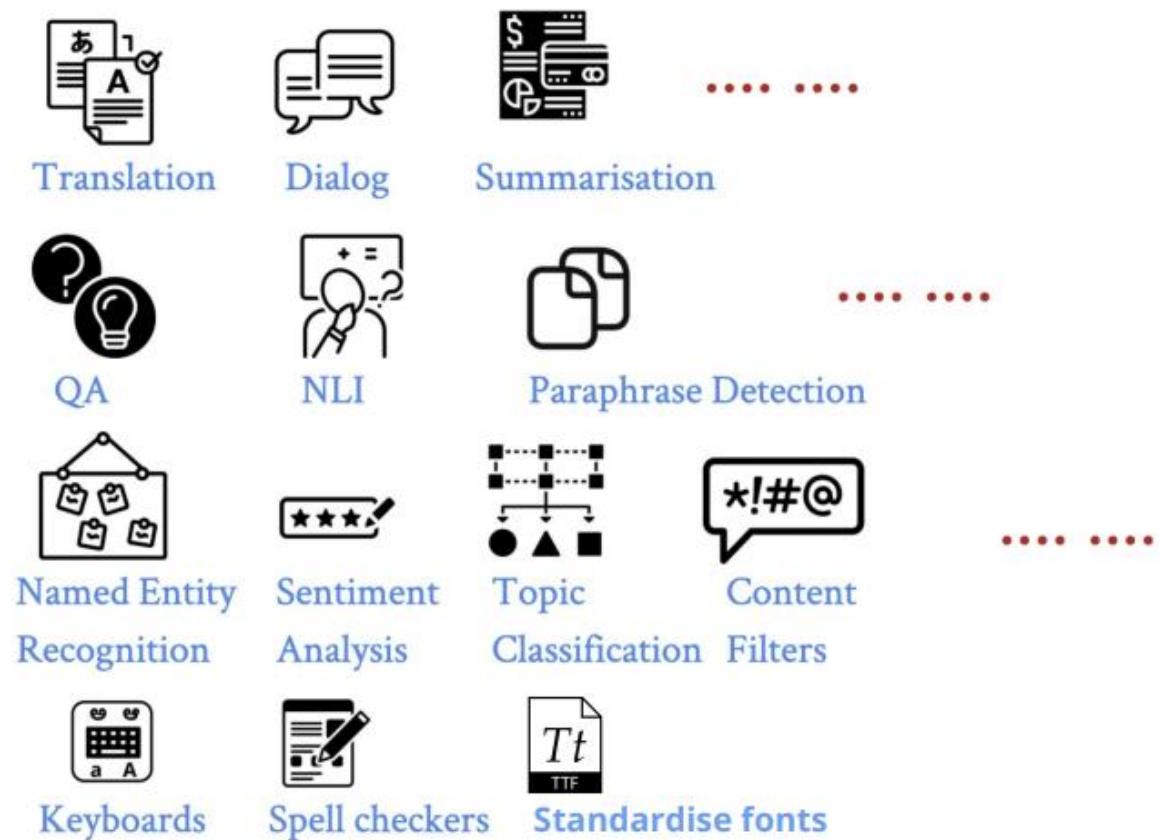
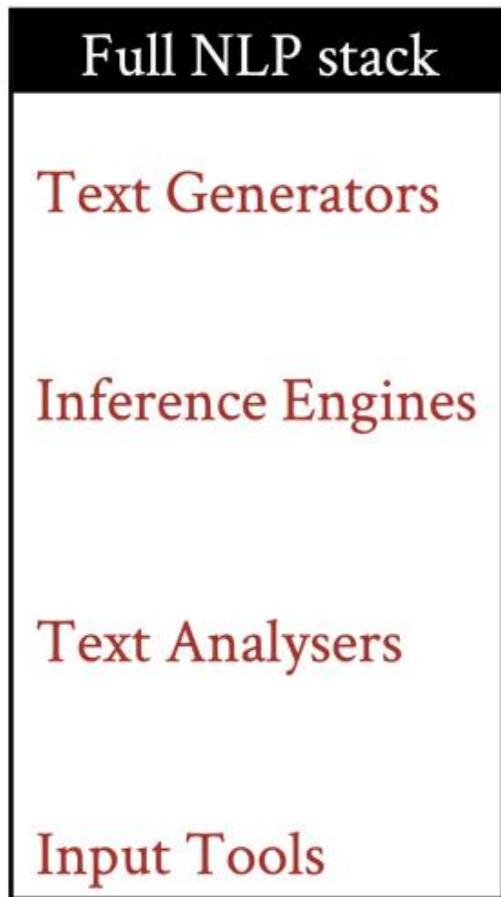
Automatic conversion of text/speech from Indian language to another language

Be the change you want to see in the world

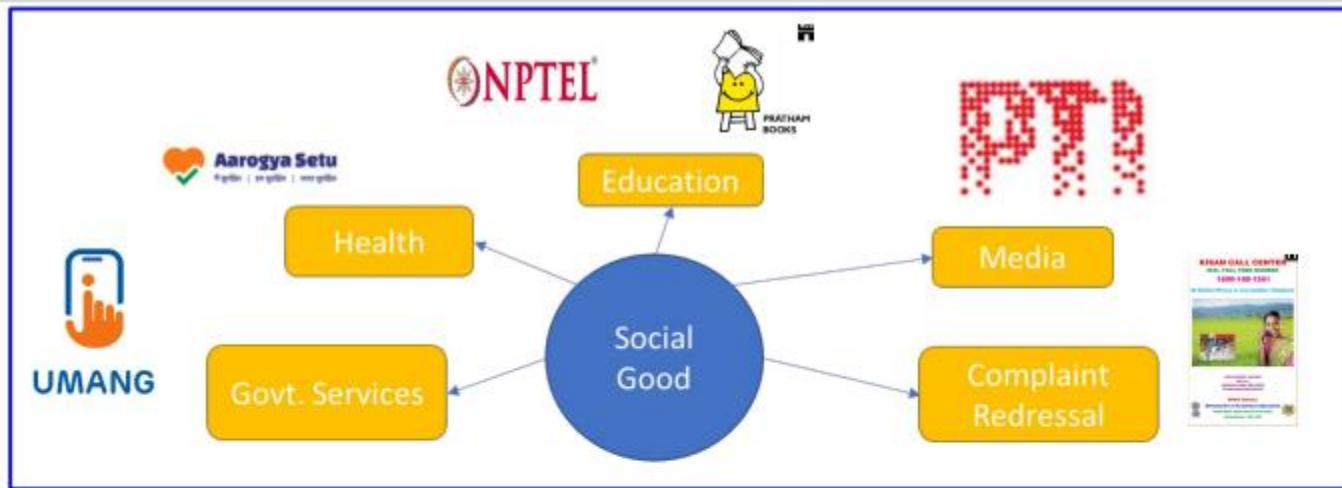
वह परिवर्तन बनो जो संसार में देखना चाहते हो



NLP for 22 constitutional languages

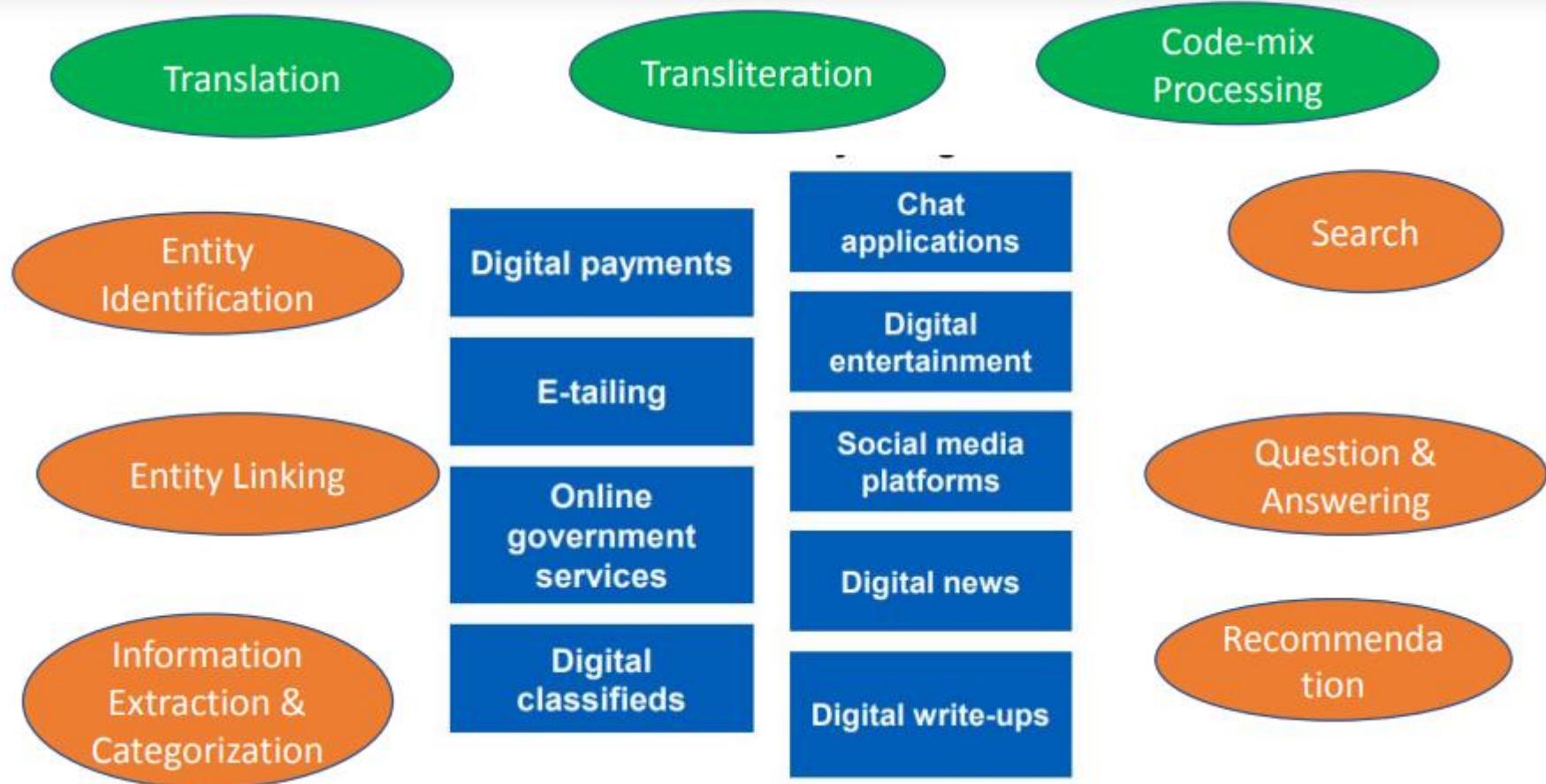


Applications- Indian Languages



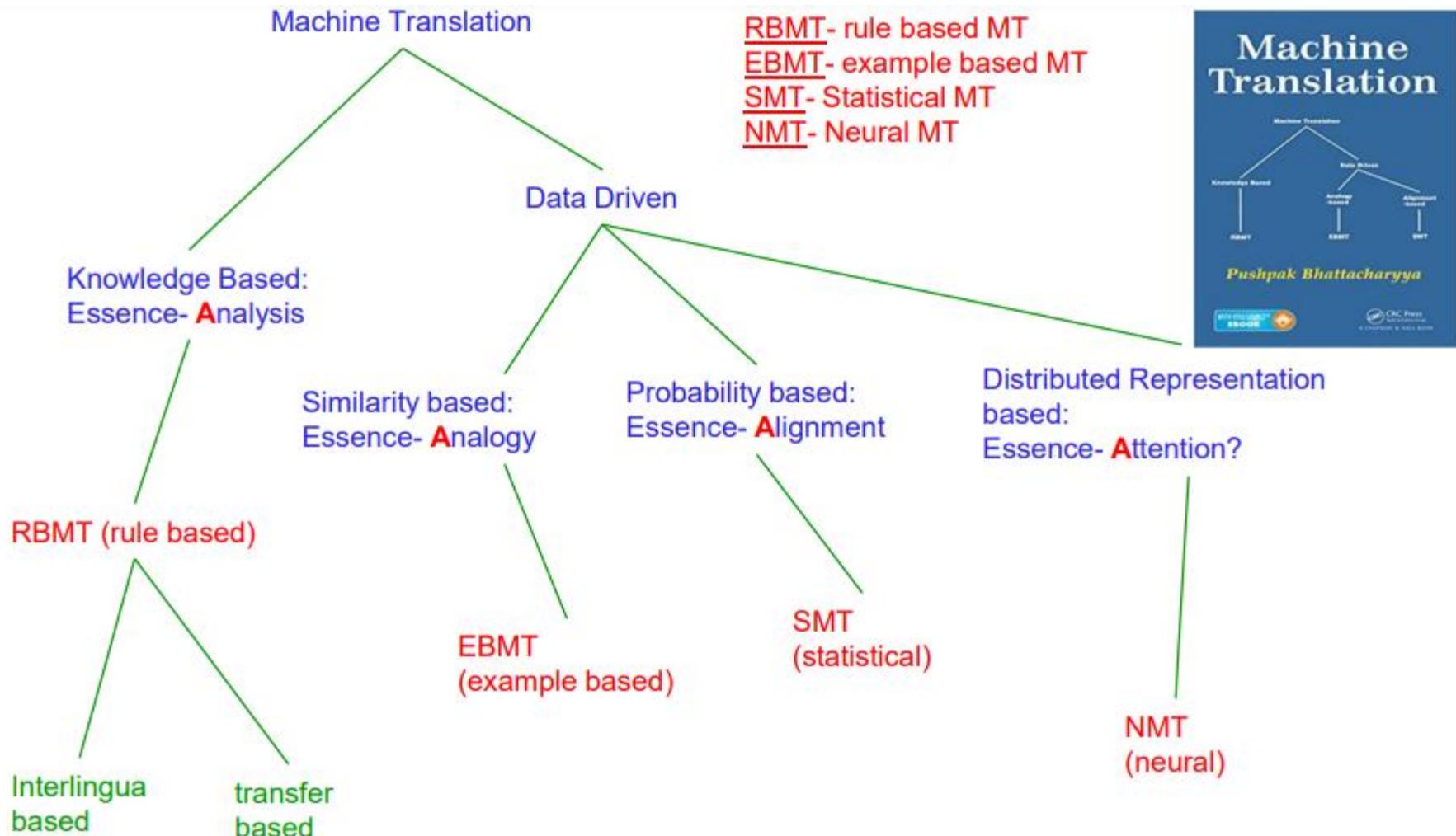
Ref: AI4Bharat Tutorial at ICON 2021, Dec 2021

Applications that require Indian Language Support



Ref: AI4Bharat Tutorial at ICON 2021, Dec 2021

Machine Translation approaches



Challenges in Indian Language translation

- **Scale and Diversity:** 22 major languages in India, written in 13 different scripts, with over 720 dialects
- **Code Mixing** (“kyo ye hesitation?”); **Gerundification** (“gaadi chalaoing”)
- **Absence of basic NLP tools and resources:** ref nlp pipeline
- **Absence of linguistic tradition for many languages**

Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga and Ranjiva Munasinghe, *Indic Language Computing*, CACM, V 62(11), November 2019.

Challenges in Indian Language translation

- **Script complexity and non-standard input mechanism:** InScript Non-optimal
- **Non-standard transliteration** (“mango”→ ‘am”, “aam”, Am”)
- **Non-standard storage:** proprietary fonts
- **Challenging language phenomena:** Compound verbs (“has padaa”), morph stacking (“gharaasamorchyaanii”)
- **Resource Scarcity**

Ambiguity in translation

English pronouns

he, she, it

Hindi pronoun

vaha

He is going to Delhi ==>

vaha dilli jaa rah**aa** hai

She is going to Delhi ==>

vaha dillii jaa rah**ii** hai

It broke ==>

vaha TuuTa ??

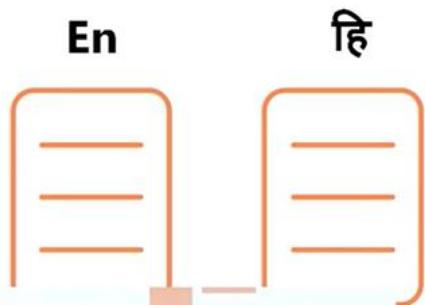
Gender Information

Information *does not always map fully from one language into another*

Conceptual worlds may be different

Modern NMT

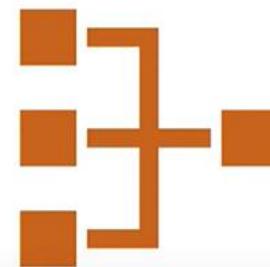
DATA



Large scale models with
innovations specific to
Indic languages



MODELS



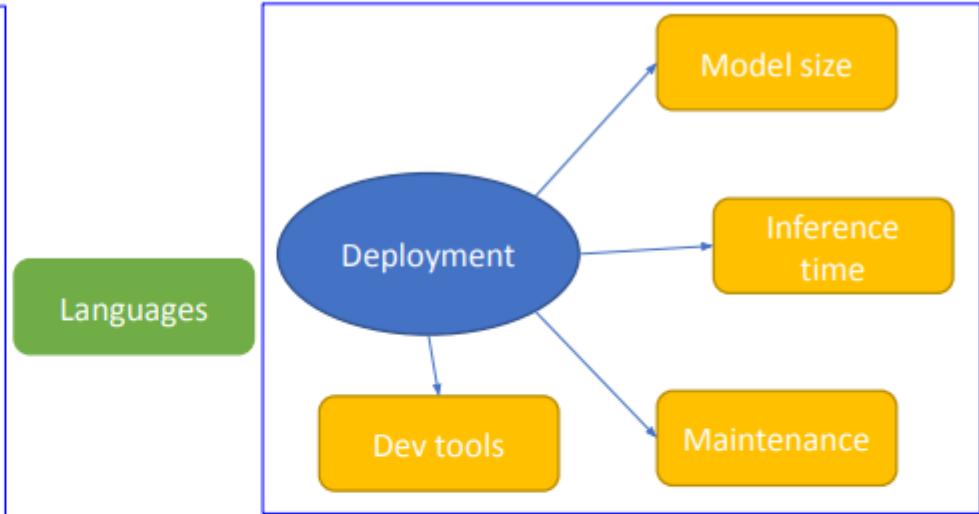
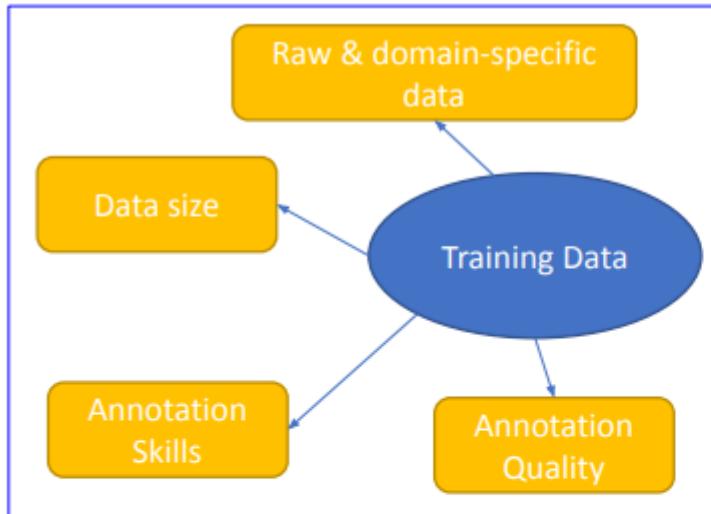
Large scale models with
innovations specific to
Indic languages

EVALUATION



Robust evaluation with
diverse benchmarks and
reliable evaluation metrics

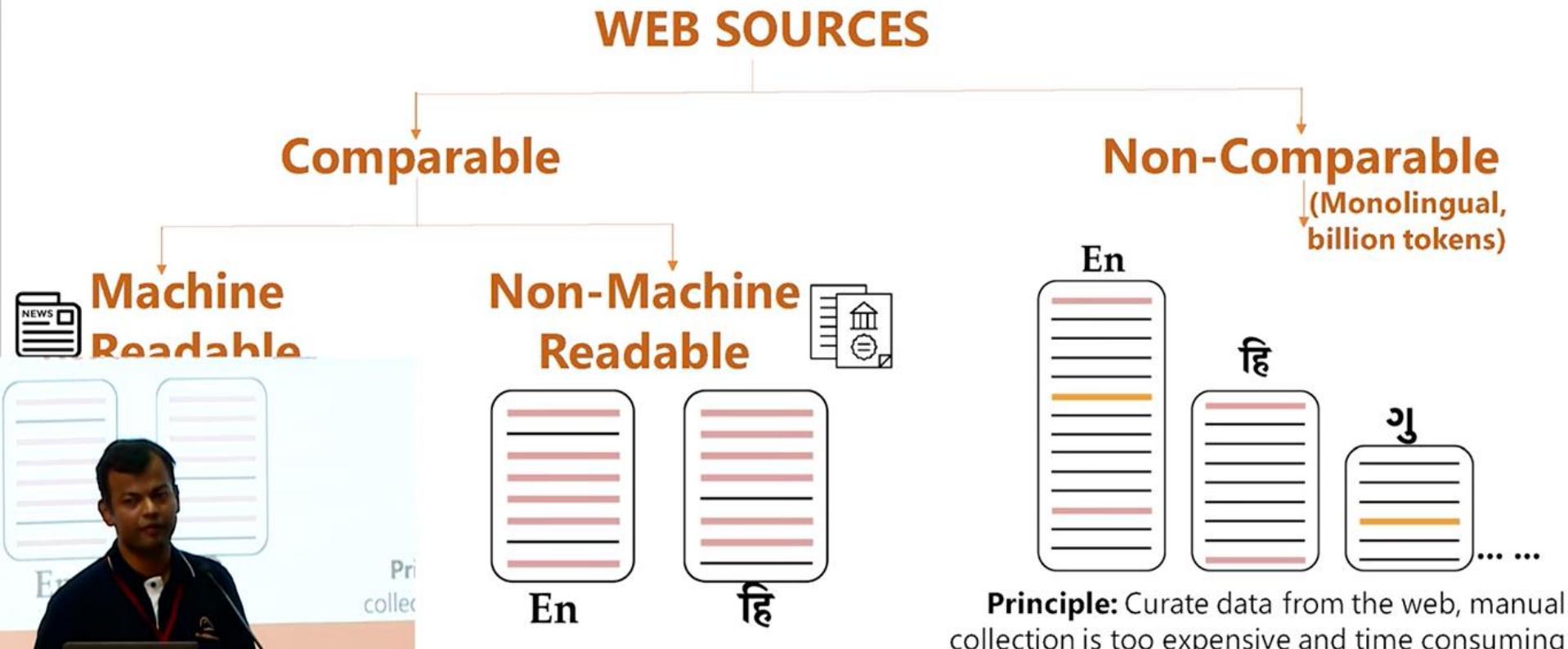
Scalability Challenges



Effort and cost increase as languages increase



How to solve data problem



How to solve data problem



En

<https://mykhel.com/>

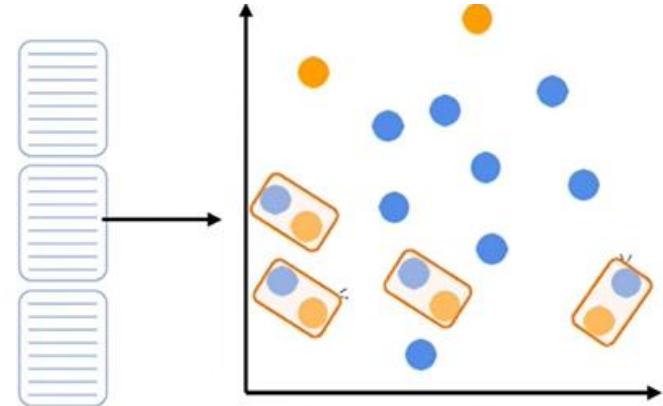


ह

indi.mykhel.com/



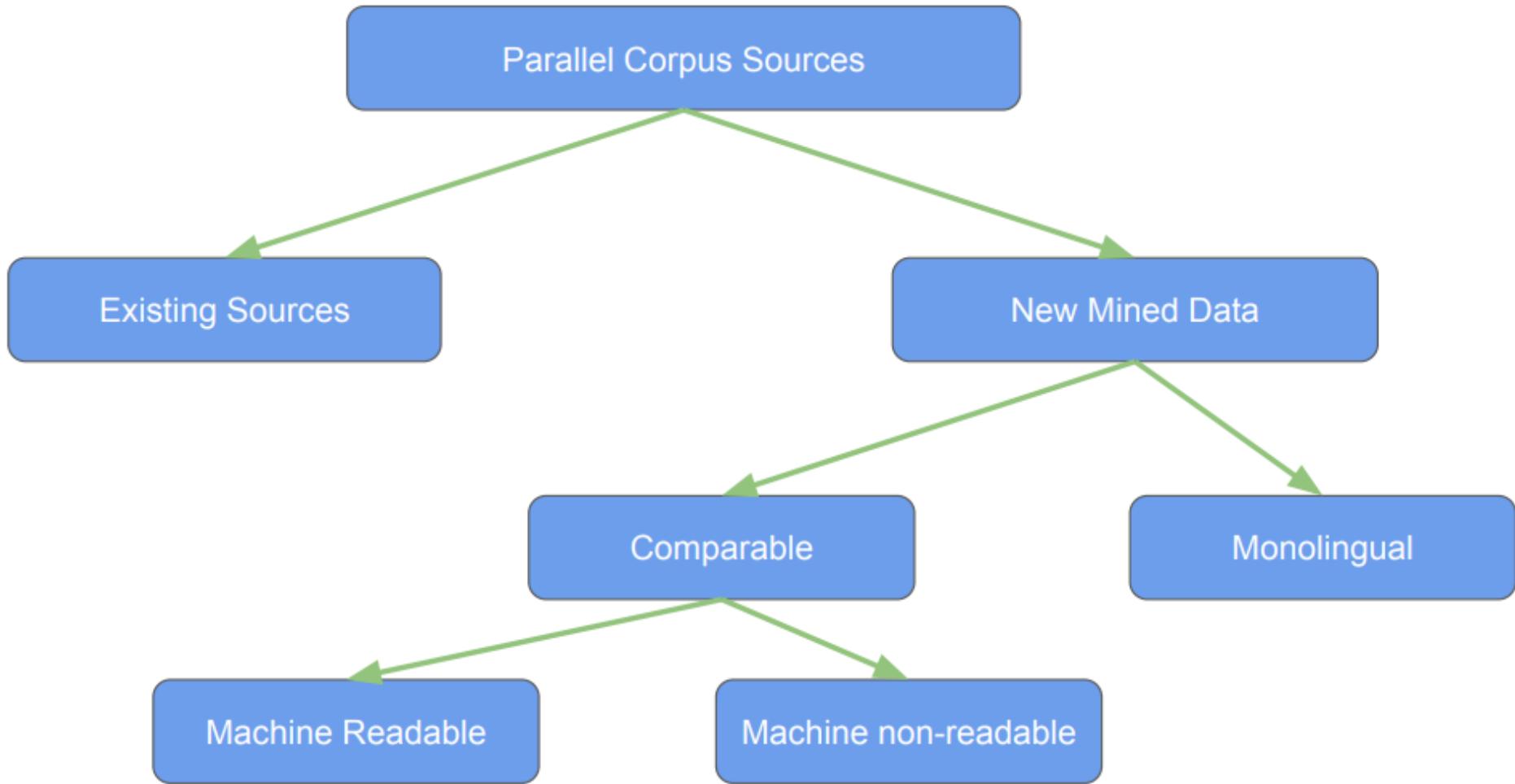
Jan 2020



Shared multilingual
space

24 such news sources considered in this
work with data from 2010 onwards

Parallel Corpora Collection for 11 Indic Languages



Mining from Machine Readable Sources

- Identified 12 websites which publish content in multiple Indian languages
DriveSpark, OneIndia, NativePlanet, MyKhel, Newsonair, DW, TimesofIndia, IndianExpress, GoodReturns, CatchNews, DD National
- Identified 2 Educational sources
NPTEL, Khan Academy

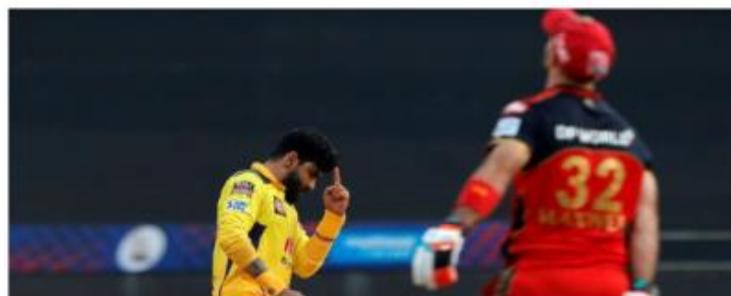
Mining from Machine Readable Sources

Home • Cricket • News • IPL 2021: RCB vs CSK: Highlights: Ravindra Jadeja show helps CSK maul RCB by 69 runs, climb at top

IPL 2021: RCB vs CSK: Highlights: Ravindra Jadeja show helps CSK maul RCB by 69 runs, climb at top

By Avinash Sharma

Updated: Sunday, April 25, 2021, 19:44 [IST]



மாதிரி: • DEL VS HYD - IN PLAY • CHE VS BAN - முடிவுமிகு போக்குவரத்து • PAK VS ZIM - முடிவுமிகு போக்குவரத்து • BAN VS SRL - முடிவுமிகு போக்குவரத்து • ZIM VS PAK - முடிவுமிகு போக்குவரத்து

செய்தி > தலைப்பு > மாநில > CSK vs RCB: மாதிரி, மாதிரி 'ஈர்' என்ற சர்வாங்கி.. மாதிரி மாநில பிளாஸ்டிக் கீட்டுப்பள்ளி

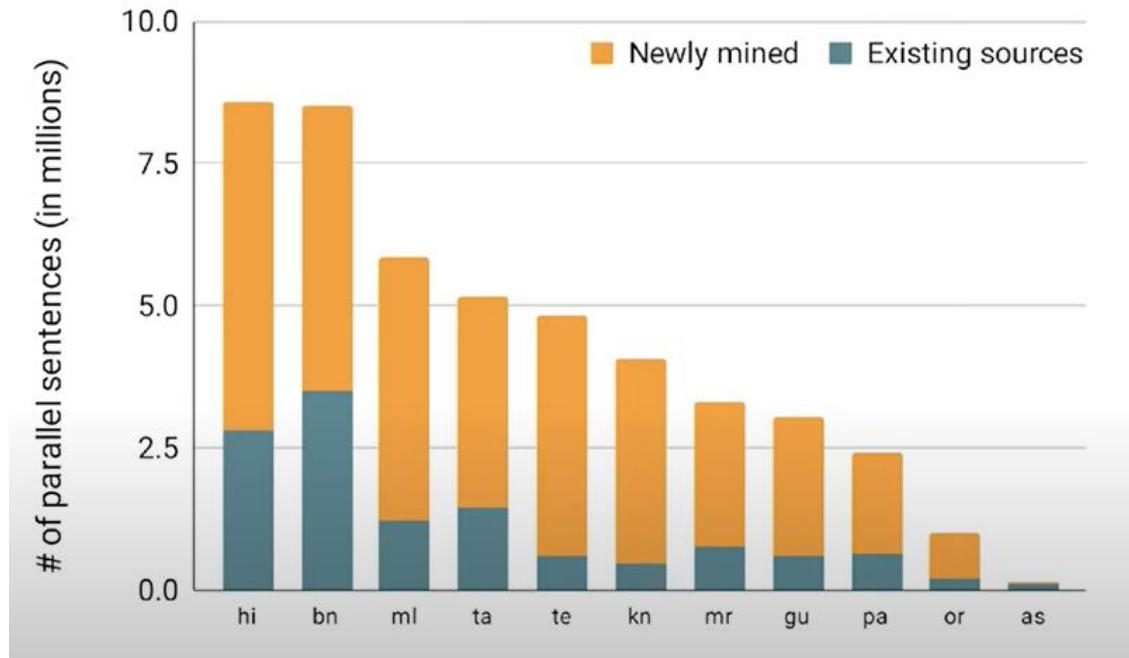
CSK vs RCB: மாதிரி, மாதிரி 'ஈர்' ஜாஜா அல்ராங்கி பீ.. பெங்க்கெழுப்பு பெங்க்கெழுப்பு! கீழீட்டு நக்கு தோலி கூடுமிகு!

By Sampath Kumar

Updated: Sunday, April 25, 2021, 19:53 [IST]



How much data collected



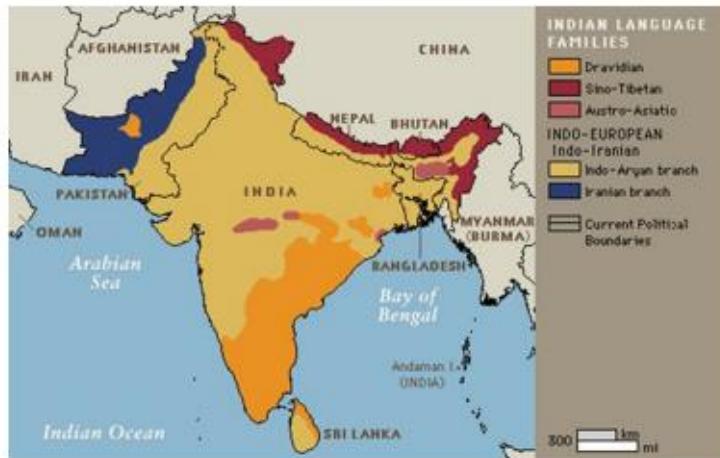
*33M parallel sentences mined from web
3 X times than earlier*

Neural machine Transliteration



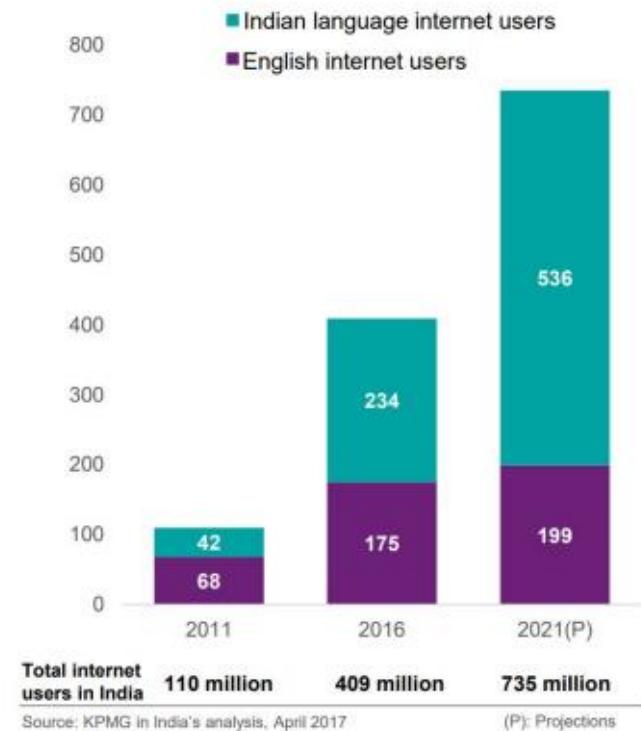
Enable Romanized typing in Indian Languages

Usage of Indian Languages



- 4 major language families
- 22 scheduled languages
- 125 million English speakers
- 8 languages in the world's top 20 languages
- 30 languages with more than 1 million speakers

Sources: Wikipedia, Census of India 2011



Internet User Base in India (in million)

Source: Indian Languages:
Defining India's Internet KPMG-Google Report 2017

The Opportunity for Indian Language NLP

Mine Datasets

Deep Learning based NLP

Representation Learning

Multilinguality

Language
Relatedness

Pre-trained
Models

*Language Agnostic
Models*

*Effective Transfer
Learning*

*Infuse linguistic and world
knowledge into models*

ML Approach

Simple Features

Bag-of-words (presence/absence)

Well-made	hit	script	lovely	boring	music
1	1	1	1	0	1

Large and sparse feature vector: size of vocabulary

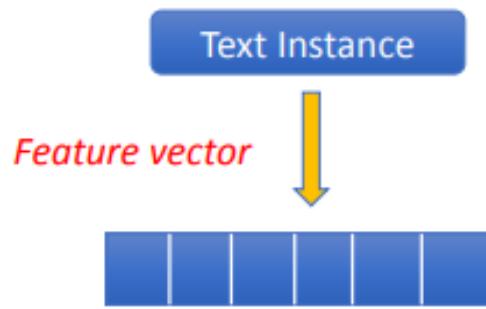
Each feature is atomic □ similarity between features, synonyms not captured

More features

- Bigrams: e.g. *lovely_script*
- Presence in [positive/negative] sentiment word list
- Negation words
- Is the sentence sarcastic (output from sarcasm classifier?)

- *These features have to be hand-crafted manually – repeat for domains and tasks*
- *Need linguistic resources like POS, lexicons, parsers for building features*
- *Can some of these features be discovered from the text in an unsupervised manner using raw corpora?*

Distributed Representations



*Can we replace the
high-dimensional, resource-heavy document
feature vector
with*

- *low-dimensional vector*
- *learnt in an unsupervised manner*
- *subsumes many linguistic features*

Distributional Hypothesis

"A word is known by the company it keeps" - Firth (1957)

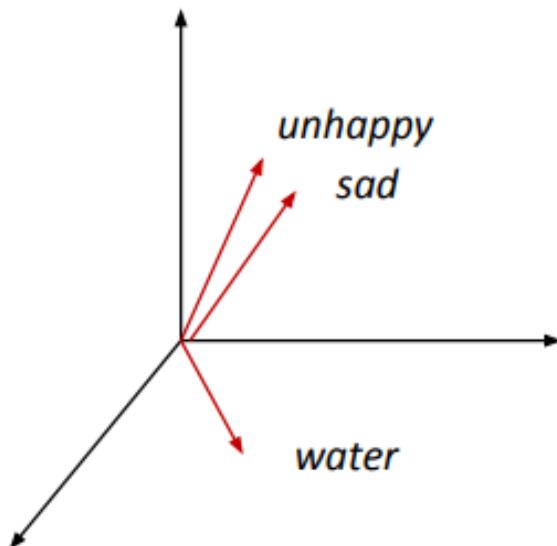
"Words that occur in similar contexts tend to have similar meanings"
- Turney and Pantel (2010)

Word2Vec

He is **unhappy** about the failure of the project

The failure of the team to successfully finish the task made him **sad**

- The distribution of the context defines the word
- Can define notion of similarity based on contextual distributions

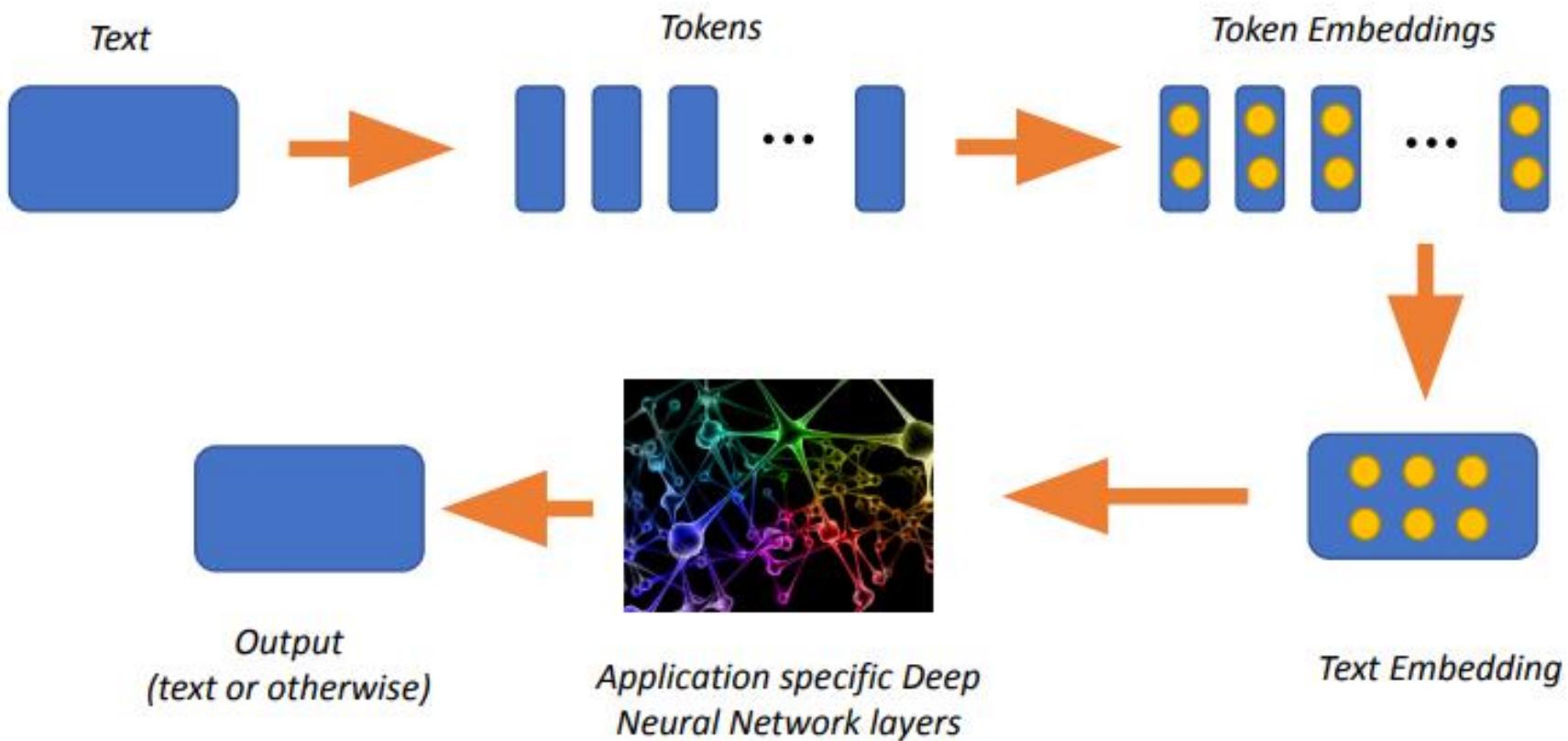


Similarity of words can be defined in terms of vector similarity: Cosine similarity, Euclidean distance, Mahalanobis distance

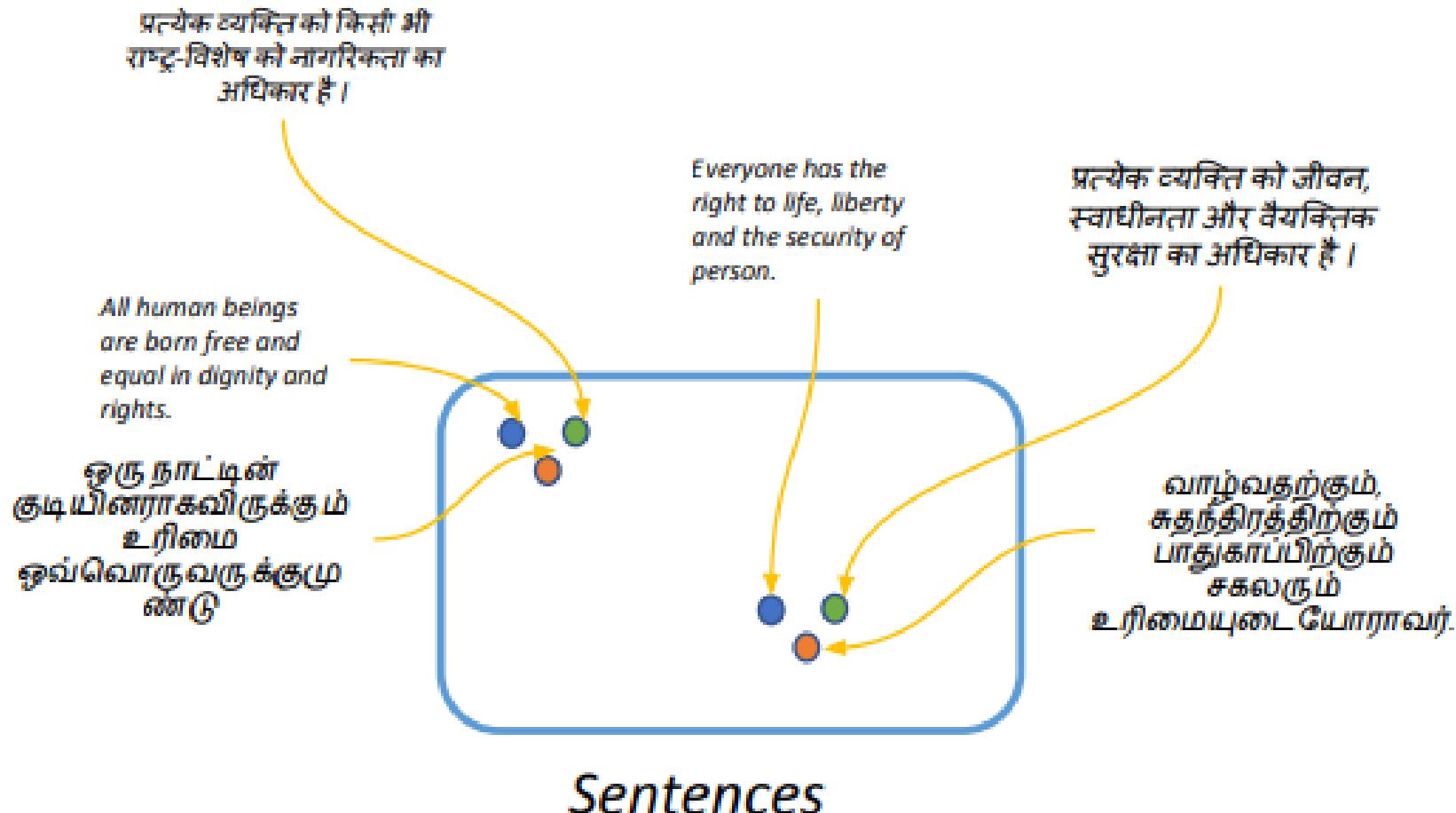
Similarity across languages

Contextual representation of words

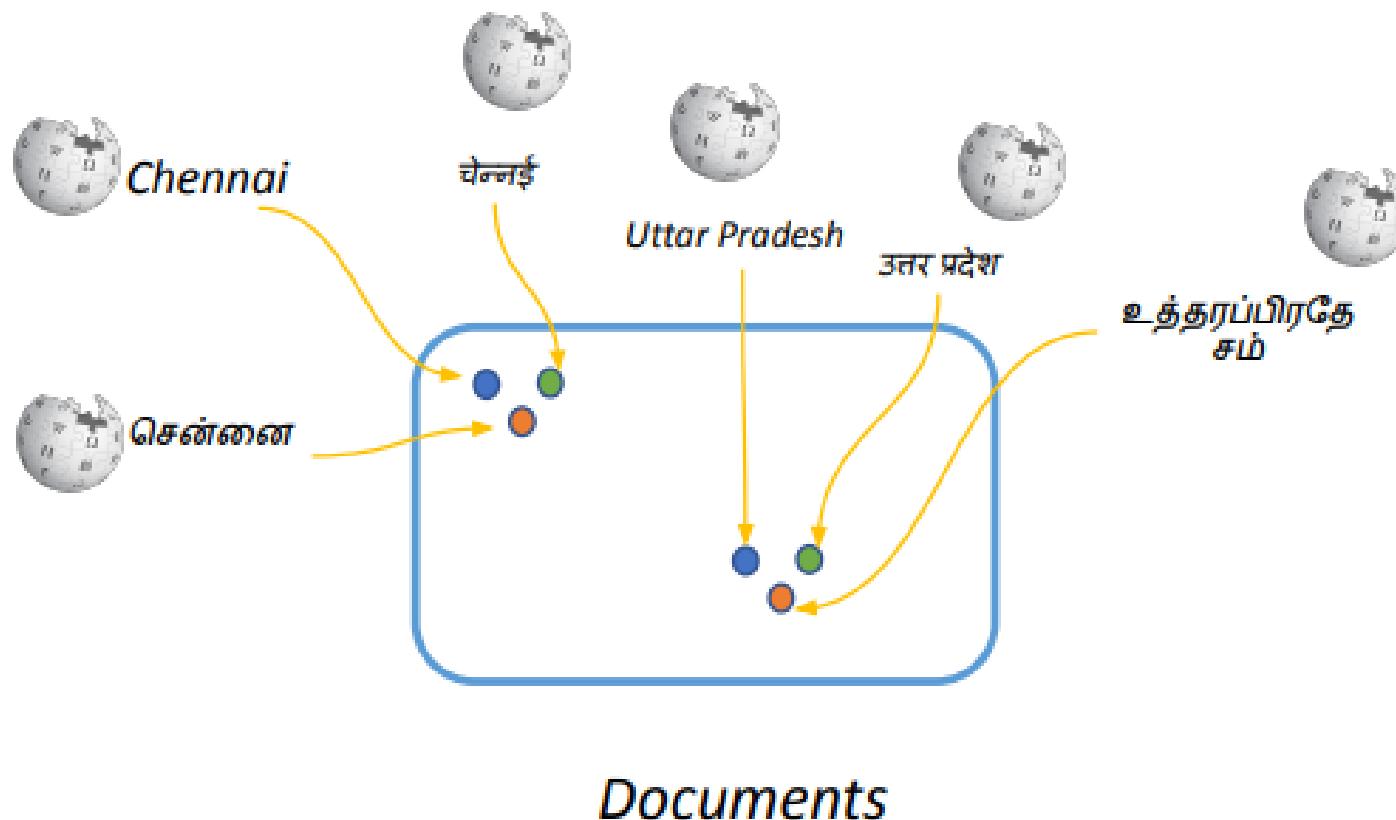
Deep Learning Pipeline



Represent semantically similar language artifacts in the same vector space

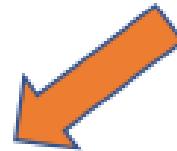


Represent semantically similar language artifacts in the same vector space



Multilinguality

Single model for multiple languages

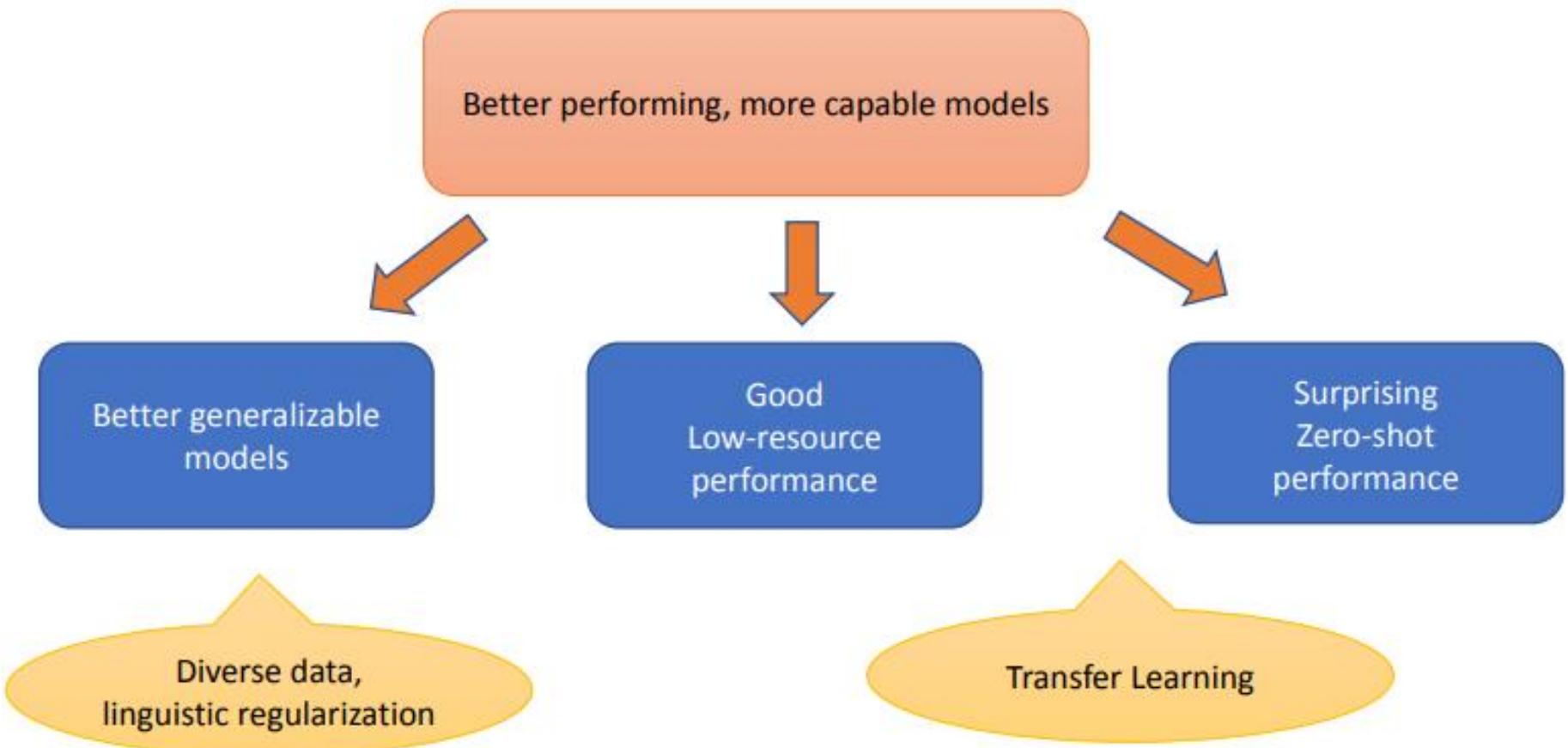


Smaller Deployment
Footprint



Easier Model
Maintenance

Multilinguality



Encoder Decoder Model

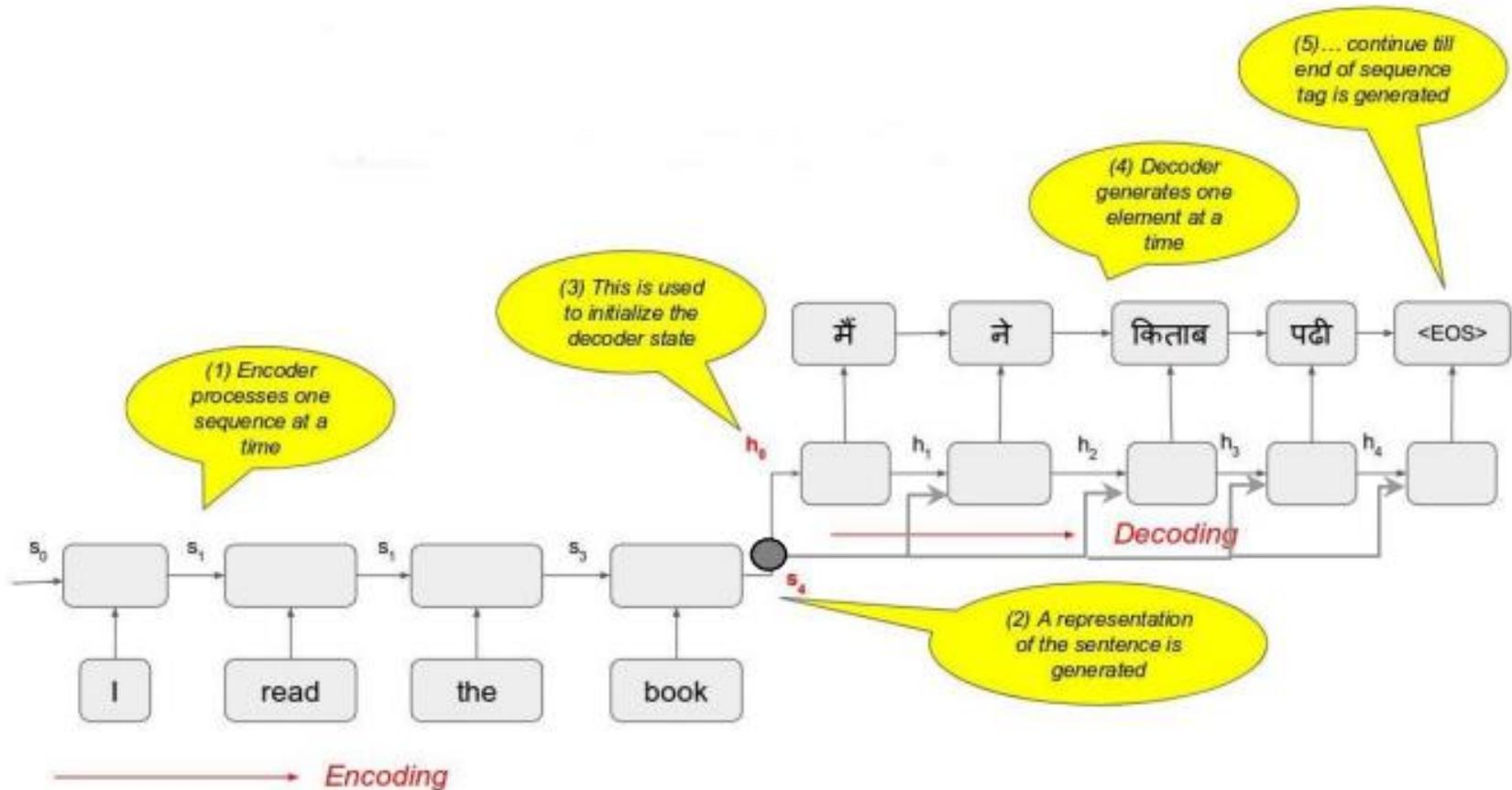
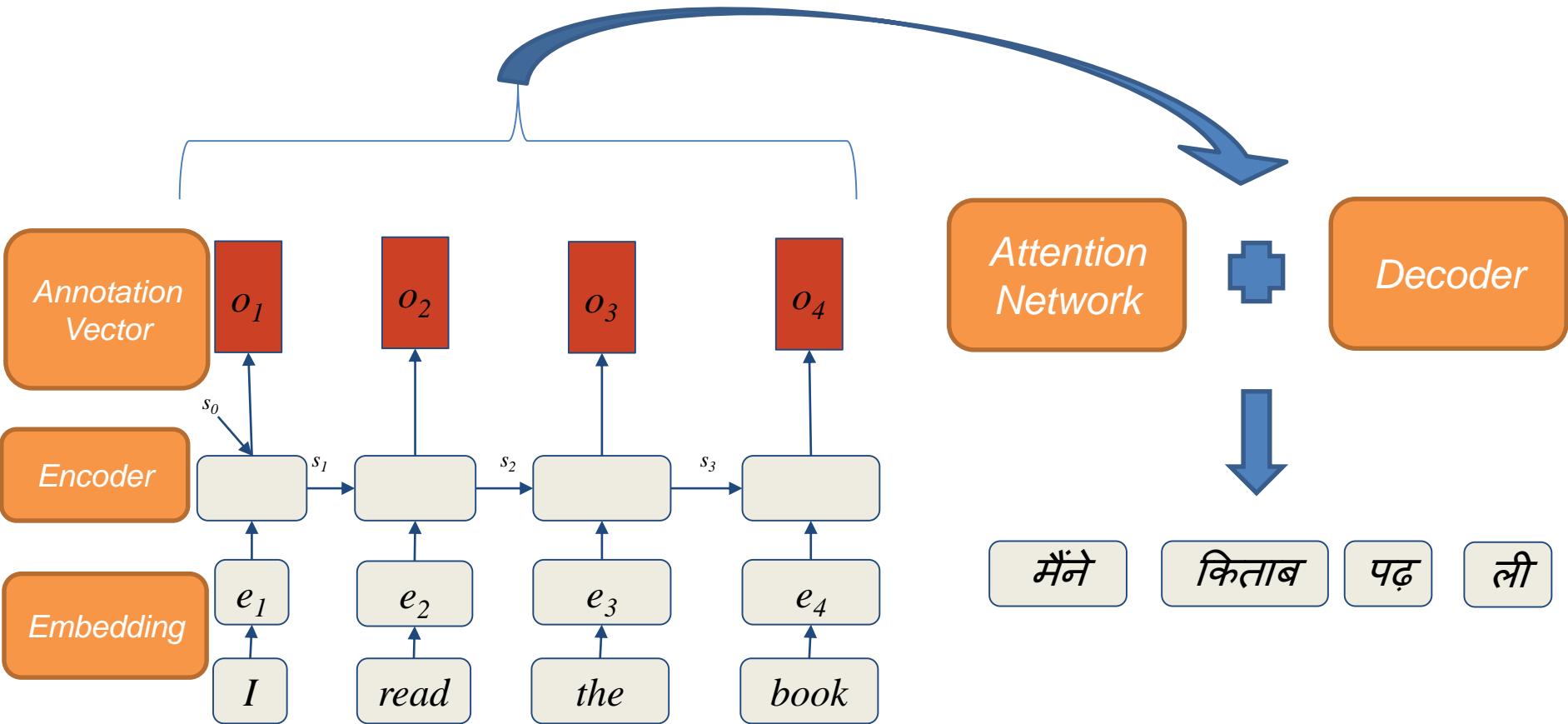


Image source- <http://www.iitp.ac.in/~shad.pcs15/data/nmt-rudra.pdf>

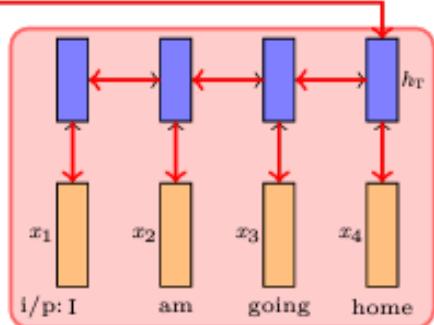
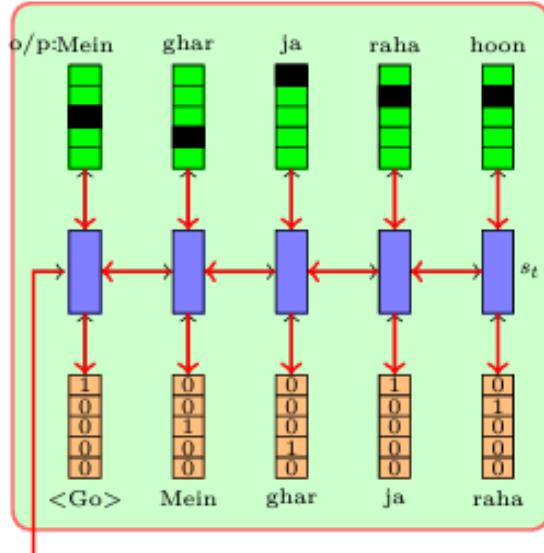
Embed - Encode - Attend - Decode Paradigm

(Bahdanau et al, 2015)



Encoder Decoder Architecture for Indic MT

o/p : Mein ghar ja raha hoon



i/p : I am going home

- **Task:** Machine translation
- **Data:** $\{x_i = \text{source}_i, y_i = \text{target}_i\}_{i=1}^N$
- **Model (Option 1):**
 - **Encoder:**

$$h_t = RNN(h_{t-1}, x_{it})$$
 - **Decoder:**

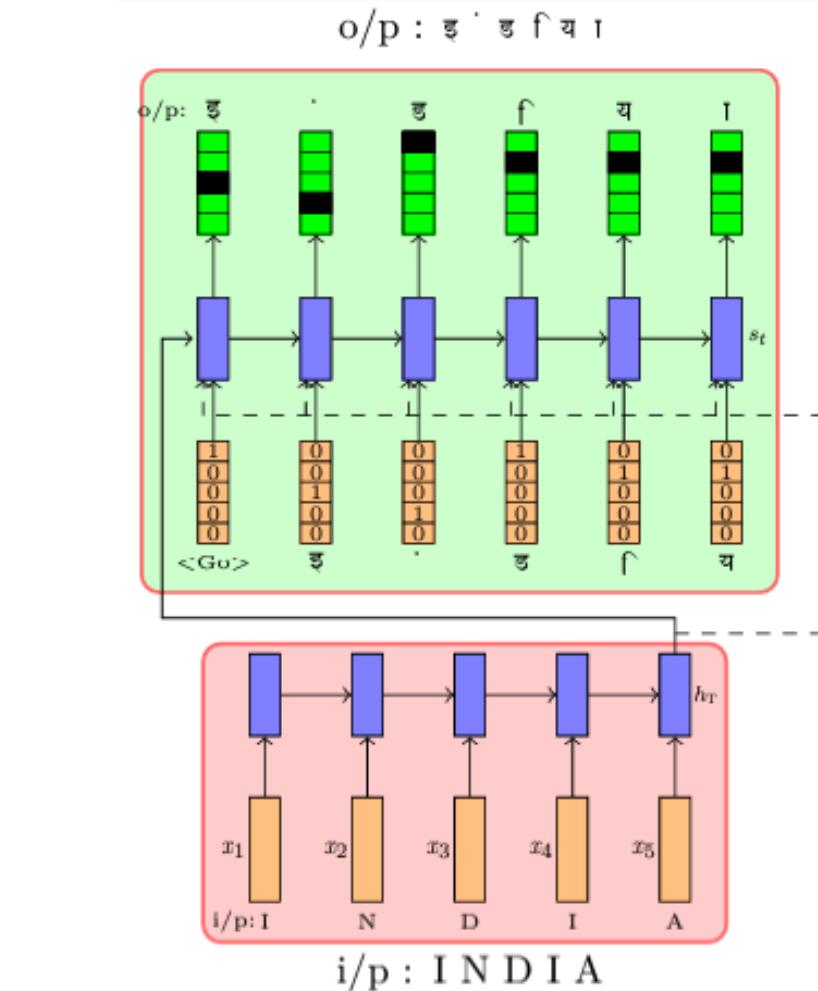
$$s_0 = h_T \quad (T \text{ is length of input})$$

$$s_t = RNN(s_{t-1}, e(\hat{y}_{t-1}))$$

$$P(y_t|y_1^{t-1}, x) = \text{softmax}(Vs_t + b)$$
- **Parameters:** $U_{dec}, V, W_{dec}, U_{enc}, W_{enc}, b$
- **Loss:**

$$\mathcal{L}(\theta) = \sum_{i=1}^T \mathcal{L}_t(\theta) = - \sum_{t=1}^T \log P(y_t = \ell_t | y_1^{t-1}, x)$$
- **Algorithm:** Gradient descent with backpropagation

Encoder Decoder Architecture for Transliteration



Transformer

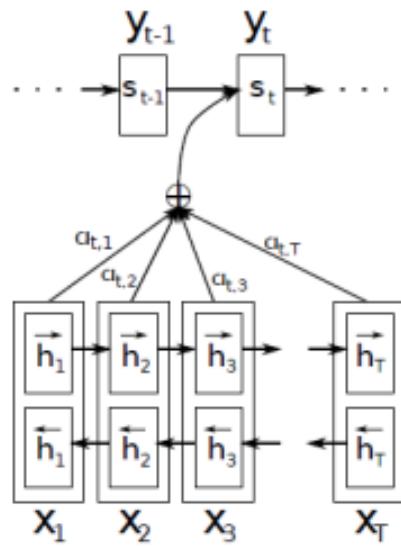
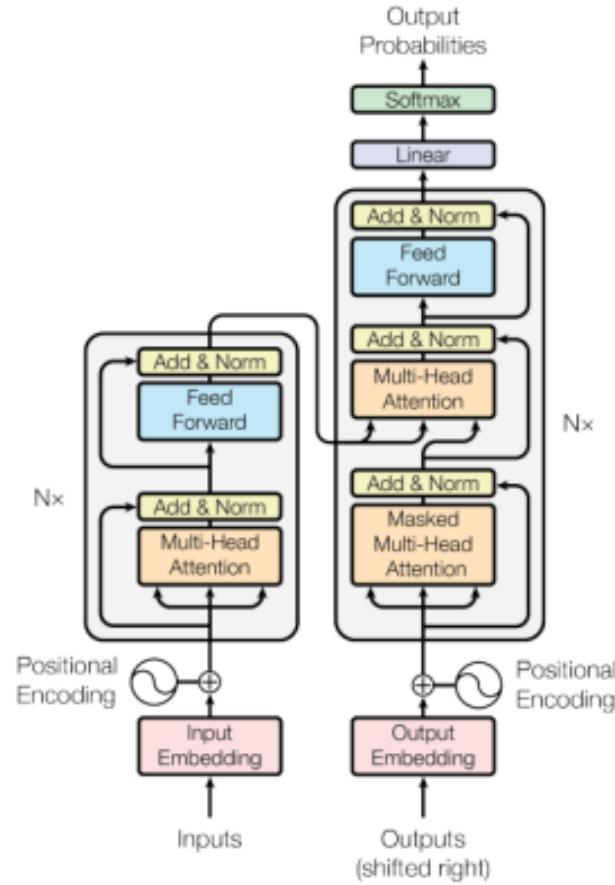


Figure 1: The graphical illustration of the proposed model trying to generate the t -th target word y_t given a source sentence (x_1, x_2, \dots, x_T) .

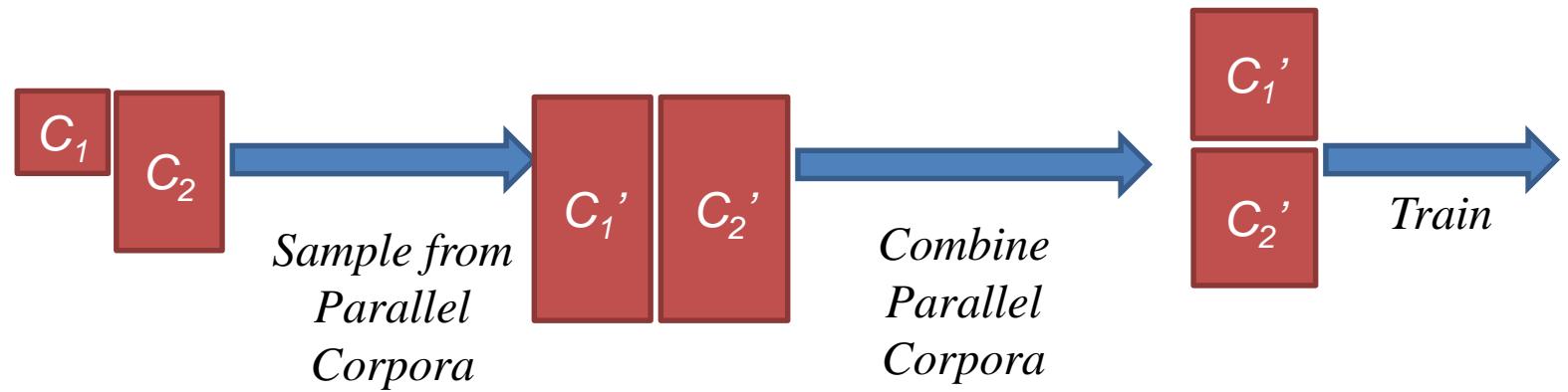
**BiLSTM encoder
decoder [3]**



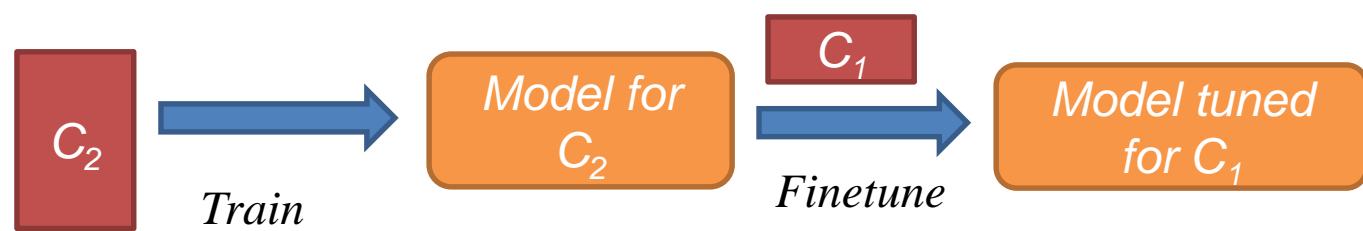
Transformer [8]

Training Multilingual NMT systems

Method 1

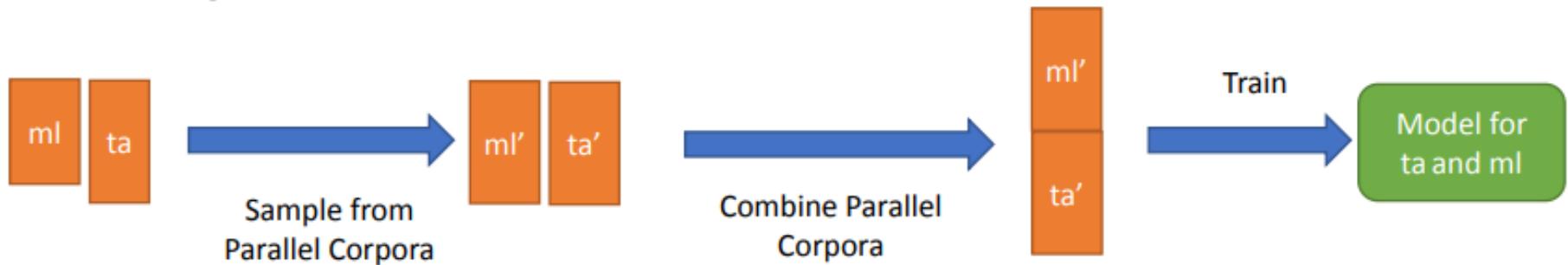


Method 2

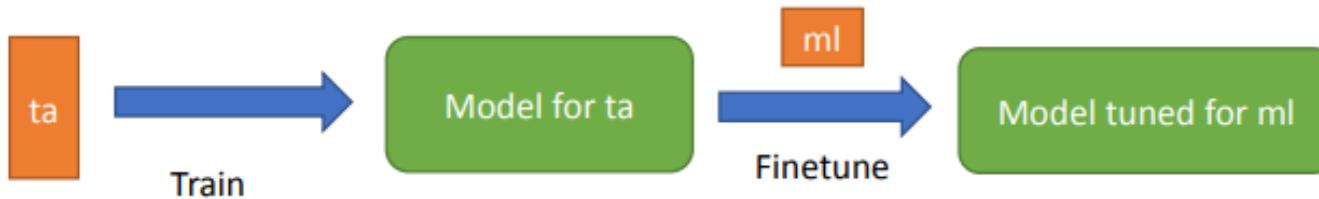


Training multilingual NMT

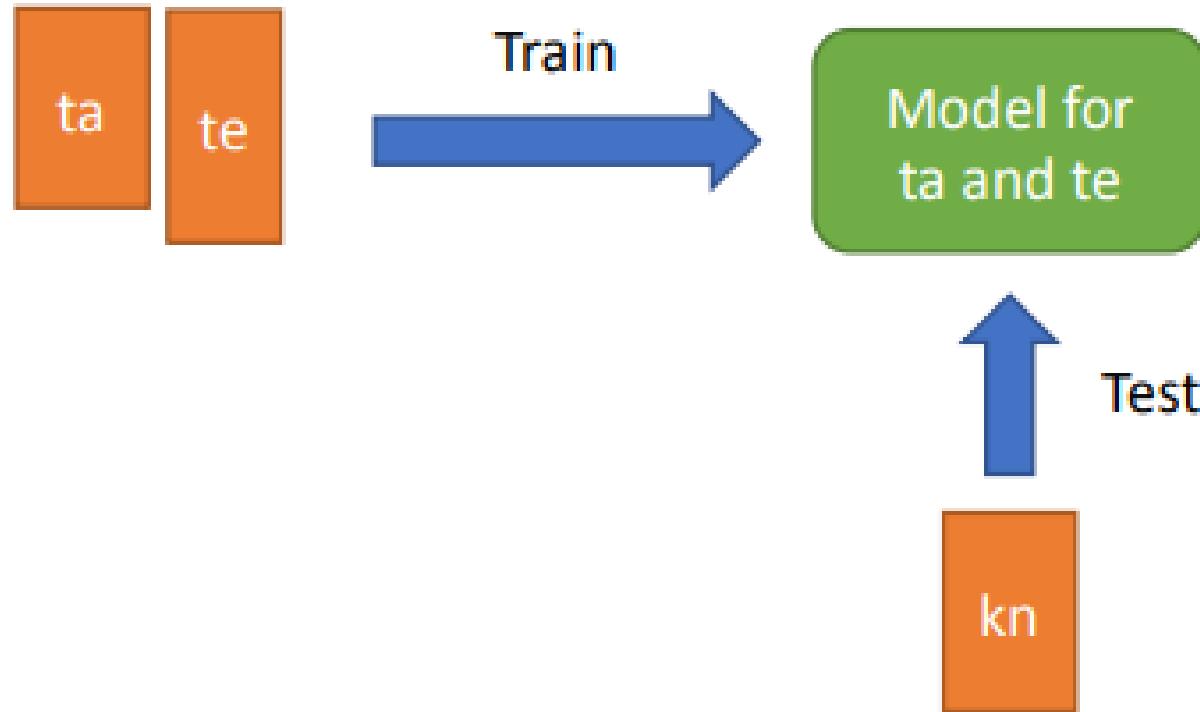
Joint Training



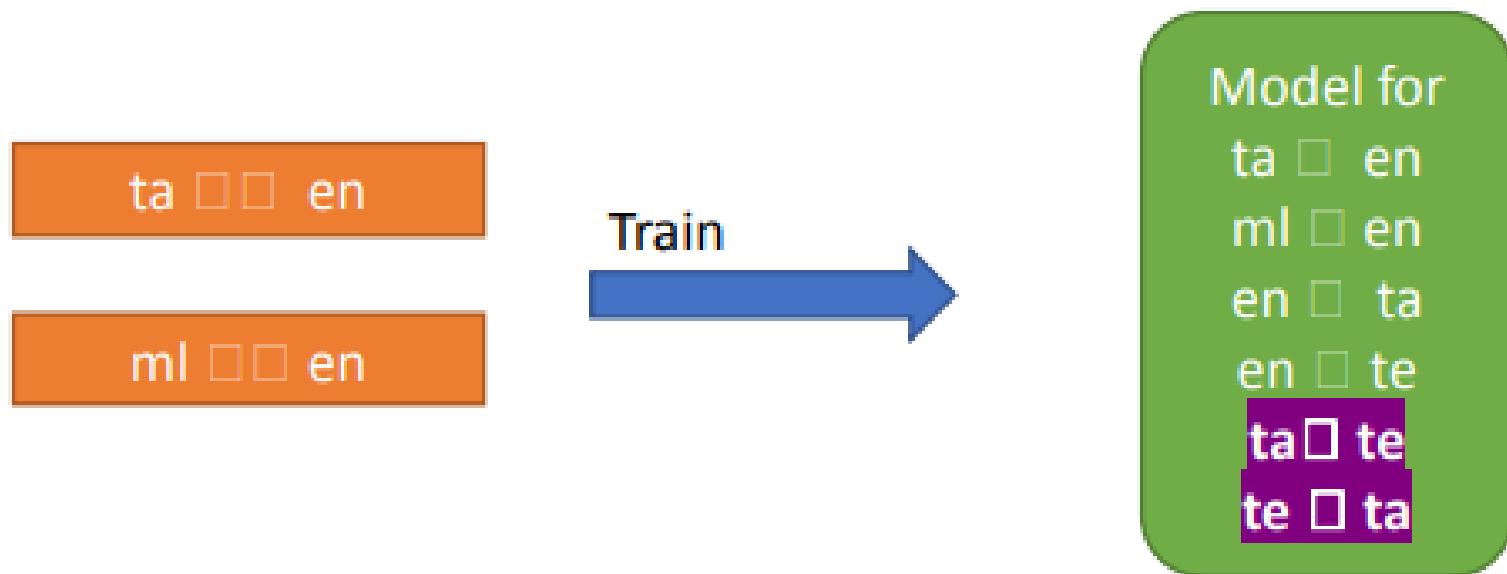
Transfer Learning



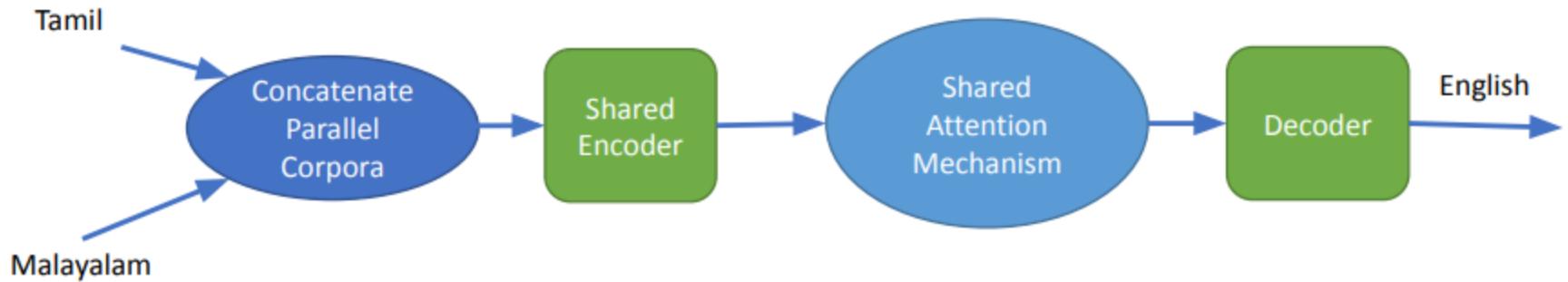
Zero shot translation to English



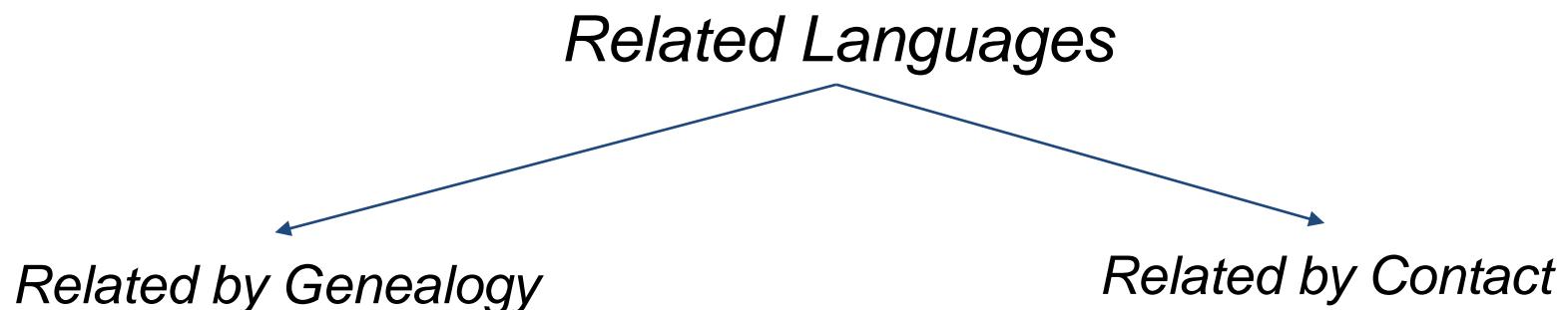
Zero shot translation between Indian Languages



Malayalam to English Translation



We want **Malayalam** □ **English** translation □ but little parallel corpus is available
We have lot of **Tamil** □ **English** parallel corpus



Language Families
Dravidian, Indo-European, Turkic

(*Jones, Rasmus, Verner, 18th & 19th centuries, Raymond ed. (2005)*)

Linguistic Areas
Indian Subcontinent, Standard Average European

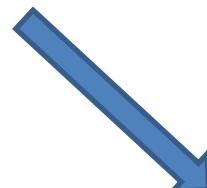
(*Trubetzkoy, 1923*)

Related languages may not belong to the same language family!

*Naturally, lot of communication between such languages
(government, social, business needs)*



Most translation requirements also involves related languages



Between related languages

Hindi-Malayalam

Marathi-Bengali

Czech-Slovak

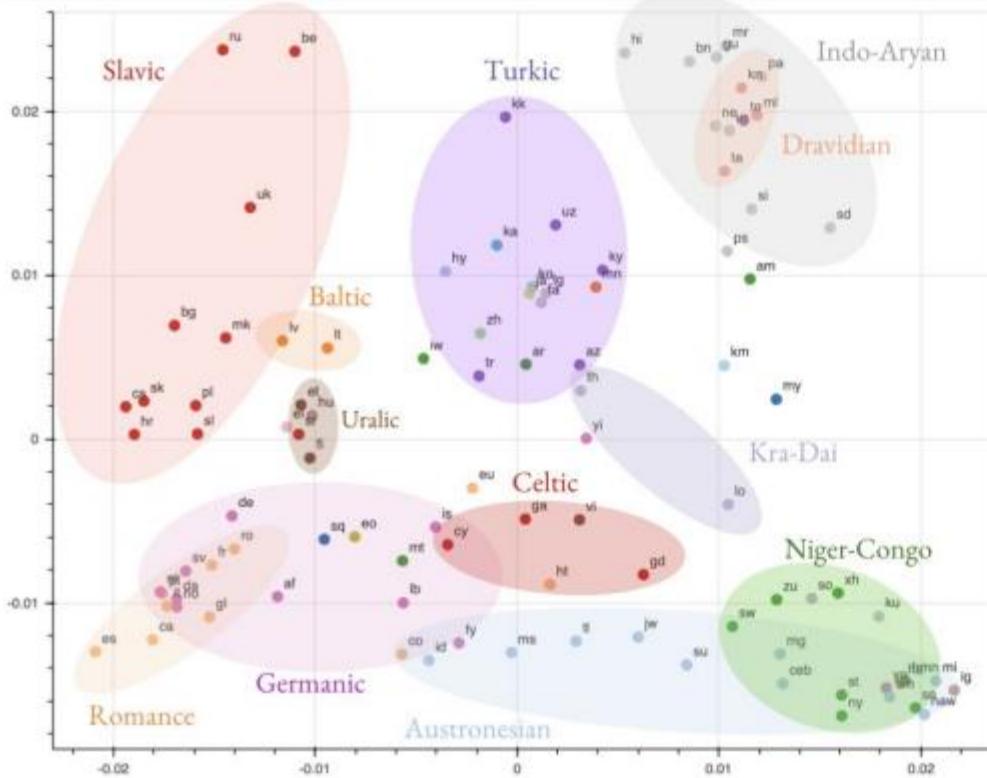
Related languages \Leftrightarrow Link languages

Kannada,Gujarati \Rightarrow English

English \Rightarrow Tamil,Telugu

*We want to be able to handle a large number of such languages
e.g. 30+ languages with a speaker population of 1 million + in the Indian subcontinent*

Transfer learning works best for related languages



(Kudungta et al, 2019) Encoder Representations cluster by language family

Transformer models are powerful enough to learn multilingual representation □ but similarity priors (natural or induced) help

Motivation for:

- *Building multilingual systems systems specific to language families*
- *Transfer learning from a related parent*

Key Similarities between related languages



भारताच्या स्वातंत्र्यदिनानिमित्त अमेरिकेतील लॉस एंजल्स शहरात कार्यक्रम
आयोजित करण्यात आला

Marathi

bhAratAcyA svAta.ntryadinAnimitta amerike tIla lOsa enjalsA shaharAtakAryakrama Ayojita karaNyAta AlA
भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एंजल्स शहरा त कार्यक्रम
आयोजित करण्यात आला

Marathi
segmented

bhAratA cvA svAta.ntrya dinA nimitta amerike tIla lOsa enjalsA shaharA ta kAryakrama Ayojita karaNyAta AlA
भारत के स्वतंत्रता दिवस के अवसर पर अमरीका के लॉस एंजल्स शहर में कार्यक्रम
आयोजित किया गया

Hindi

bhArata ke svata.ntratA divasa ke avasara para amarIkA ke losa enjalsA shahara me.n kAryakrama Ayojita kiyA gayA

Lexical: share significant vocabulary (cognates & loanwords)

Morphological: correspondence between suffixes/post-positions

Syntactic: share the same basic word order

Key Similarities between related languages



Brahmi-derived Indic scripts are orthographically similar

Devanagari	अ आ इ ई उ ऊ ऋ ल्ह एँ एॅ एै ए॒ ओ ओ॑ औ औ॒ क ख ग घ ङ च छ ज झ
Bengali	অ আ ই ঈ উ ঊ র ল্হ এঁ এ৅ এৈ এ৒ ও ও৑ ক খ গ ঘ ঙ চ ছ জ ব এট ঠ ণ ত ষ
Gurmukhi	ਅ ਆ ਇ ਈ ਉ ਊ ਰ ਲ੍ਹ ਏਂ ਏਅ ਏਈ ਕ ਖ ਗ ਘ ਙ ਚ ਛ ਜ ਵ ਏਟ ਠ ਣ ਤ ਷
Gujarati	અ આઈ ઈ ઉ ઊ રી એ એઈ એઉ ઓ ઓા ઓિ ક ખ ગ ઘ ઙ ચ છ જ બ એટ ઠ
Oriya	ଅ ଆ ଇ ଈ ଉ ଊ ର ଈ ଏ ଏପ୍ ଓ ଏପ୍ କ ଖ ଗ ଘ ଙ ଚ ଛ ଜ ବ ଏଟ ଠ
Tamil	அ ஆ இ ஈ உ ஊ எ ஐ ஒ ஒ ஒளா க வ ச ஐ ஞ ட ன த ந
Telugu	అ ఆ ఇ శా ఈ ఉ ము ఏ ఎ ఏ ప ఒ క ఫ గ ఫు జ చ ఛ రు
Kannada	ಅ ಆ ಇ ಶಾ ಉ ಮು ಏ ಎ ಏ ಪ ಒ ಕೆ ಫಿ ಗ ಫು ಜ ಚ ಛ ರು ಏ
Malayalam	അ ആ ഇ ഉ മു ഏ ഉ മു മു എ ഫി ഫു ഫു ജ ച ഛ രു ഏ

- Largely overlapping character set, but the visual rendering differs
- *highly overlapping phoneme sets*
- Highly consistent grapheme-to-phoneme mapping

Script Conversion

- Read any script in any script
- Unicode standard enables **consistent script conversion with a single rule**

unicode_codepoint(char) - Unicode_range_start(L₁) + Unicode_range_start(L₂)

	DAB	DAG	DAH	DAB	DAC	DAD	DAE
0	અૈ	ં	ર	રુ	િ	ંસ	ાં
1	ઓ	ન	ા		ાં	ાં	ાં
2	ા	ણ	લ	ા	ા	ા	ા
3	ાં	ઓ	ાં	ાં	ાં	ાં	ાં
4	ાં	ાં	ાં	ાં	ાં	ાં	ાં
5	અ	ં	થ	ં	ા	ા	ા

	098	099	09A	09B	09C	09D	09E
0	ଏ	ଇ	ଠ	ର	ି	ା	ା
1	ୟ	ୟ	୭	୭	୭	୭	୭
2	୧	୧	୮	୮	୮	୮	୮
3	୭	୭	୭	୭	୭	୭	୭
4	୭	୭	୭	୭	୭	୭	୭
5	୭	୭	୭	୭	୭	୭	୭

કેરલા

kerala

કેરલા

કેરલા

As a developer, you can read text in a script you understand

Only a single mapping needed for Romanization too

Indian Language Speech sound Label set

(Samudravijaya & Murthy, 2012)

Lexical Similarity

(Words having similar form and meaning)

- *Cognates*

a common etymological origin

roTI (hi)	roTIA (pa)	bread
bhai (hi)	bhAU (mr)	brother

- *Named Entities*

do not change across languages

mu.mbal (hi)	mu.mbal	mu.mbal (pa)
keral (hi)	k.eraLA (ml)	keraL (mr)

- *Loan Words*

borrowed without translation

matsya (sa)	matsyalu	fish
	(te)	
pazha.m	phala (hi)	fruit
(ta)		

- *Fixed Expressions/Idioms*

MWE with non-compositional semantics

dAla mA mEnKAchukAIA	(hi)
honA	Something fishy
dALa mA kAIka kALu hovu	(gu)

Enables sharing of data across languages

Multilinguality

Representation Learning

*Automatic Feature Extraction
Continuous Space Representation
Numerical Optimization at disposal*

Multilingual learning

*Transfer Learning
Better generalizability across languages*

Supervised data not sufficient

*How do we understand linguistics similarities □
synonymy, parts-of-speech, word categories, analogies*

How do we know if the sentence is grammatically correct?

How do we know if the sentence makes sense?

These capabilities are important for generalization

Pre-trained Models

Task-independent models that know about language

NN Models

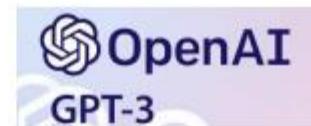
Word Embeddings

fastText

Encoder Language Model for NLU



Decoder Language Model for NLG



Encoder-decoder Language Model for NLU+NLG



 *Multilinguality*

MUSE

mBERT

mBART

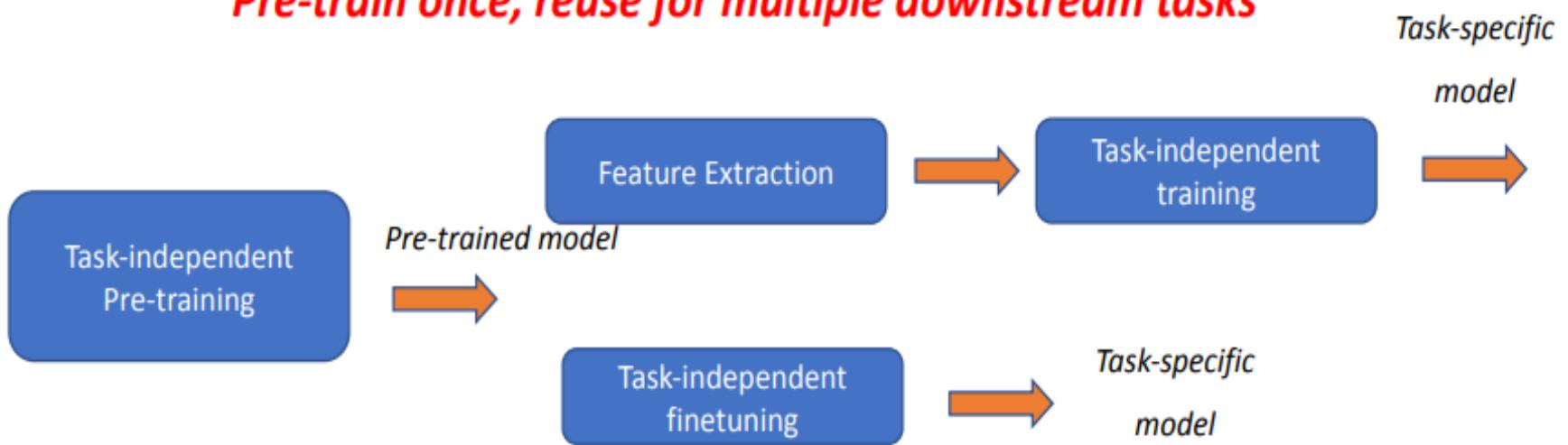
Trained on a large amount of raw text corpora with unsupervised objectives

Language models are

- *computationally intensive to train*
- *trained on a large amount of raw text corpora*
- *giant models*

NN Models

Pre-train once, reuse for multiple downstream tasks



Only task-specific training: less data & less computation

Language understanding for tasks like sentiment analysis, question answering, paraphrase detection

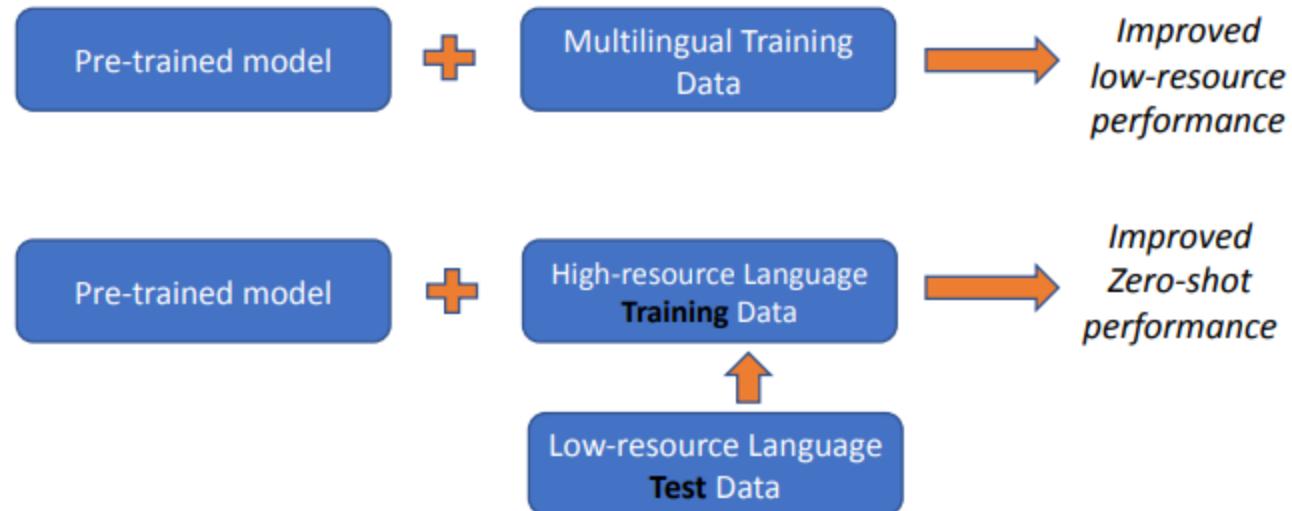
Language modeling & Language generation for tasks like summarization, ASR, question generation

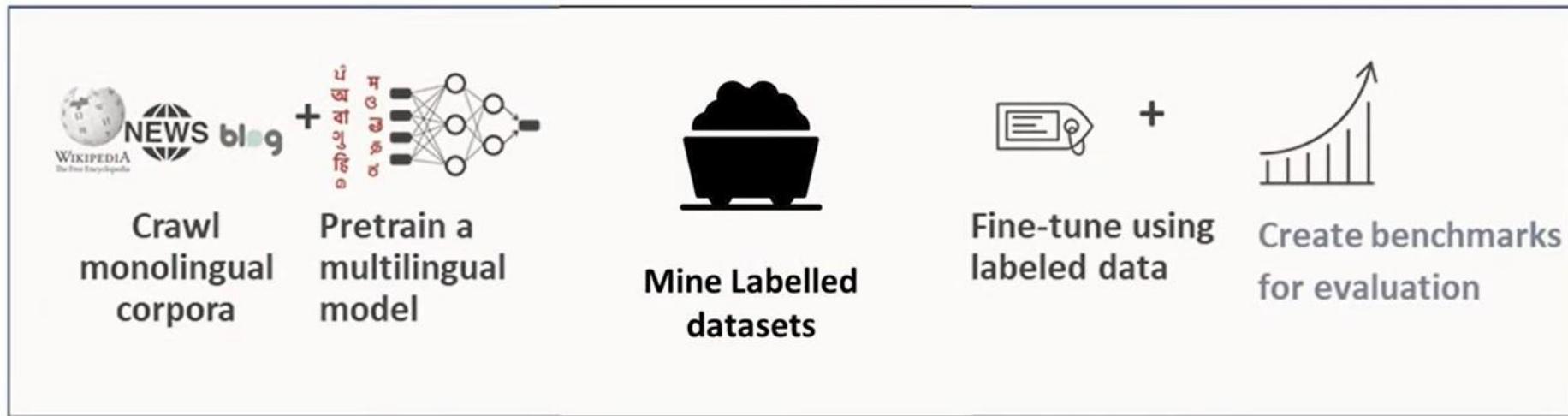
Multilinguality and NN Models

Multi-linguality and Pre-training are complementary

Language-family specific pre-trained model

- Compact pre-trained models
- Utilize language relatedness
- Better data representation





IndicBERT
IndicCorp

IndicBART

IndicGLUE
Naampadam

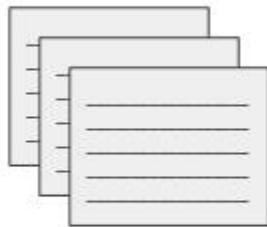
Indic NLG
Benchmark

IndicNER

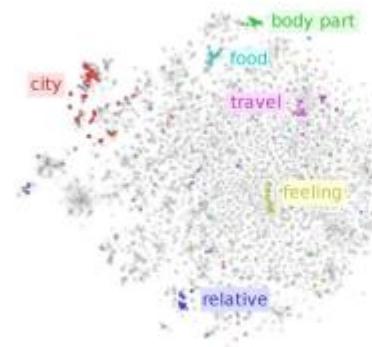
IndicNLPSuite



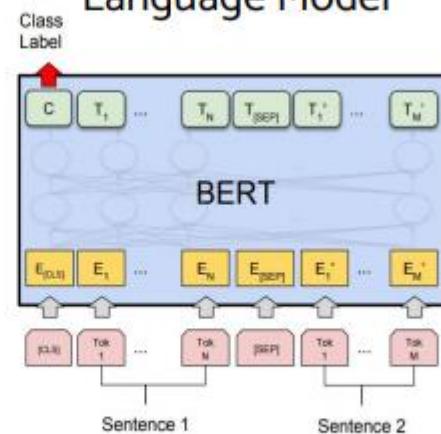
Monolingual Corpora



Embeddings



Language Model



NLU Benchmark



IndicCorp

IndicFT

IndicBERT

IndicGLUE



Large scale
Monolingual
Corpora



IndicCorp

450M
Sents.



Evaluation
benchmarks



IndicGLUE

11
Tasks

Coming soon
for IN-22



Multilingual
Language
Model for IN-
22



IndicBERT

18M
Parameter
Model

Models and Benchmark datasets

Models

- IndicBERT
- IndicBART
- n-gram LM
- IndicWav2Vec
- MT Models

IndicCorp is a central resource

Mined Datasets

- Parallel Translation Corpus
- Parallel Transliteration Corpus
- NER Corpus
- Text Classification
- Language Generation

Benchmark Datasets

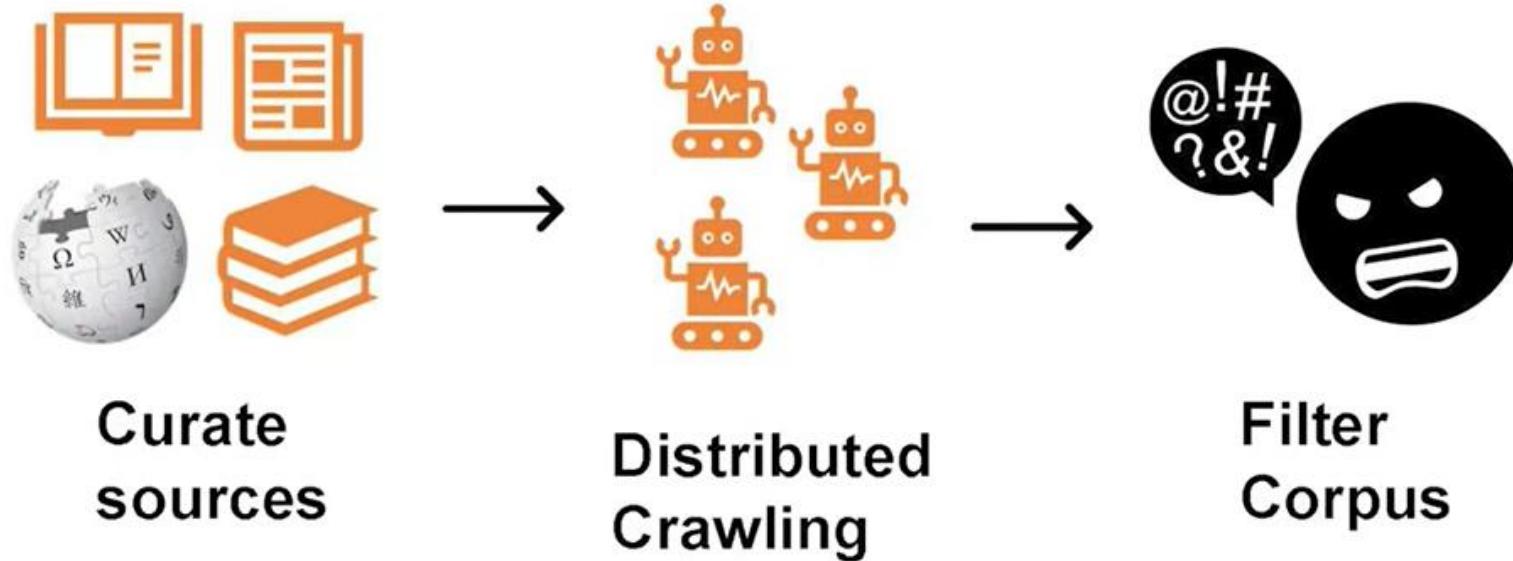
Processing Web content

Processing HTML Pages to Get Sentences

<https://github.com/AI4Bharat/webcorpus>



Monolingual Corpora collection



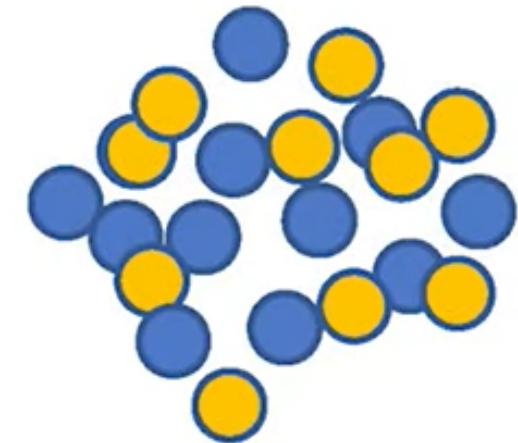
<https://github.com/AI4Bharat/webcorpus>

Multilingual Word Embedding

மரத்தாலன் (wooden)

மரத்தால் (tree) + ஆன
(making)

Complex tense, verb
embedded into a
single word



Indic FastText

IndicGlue - Benchmarks

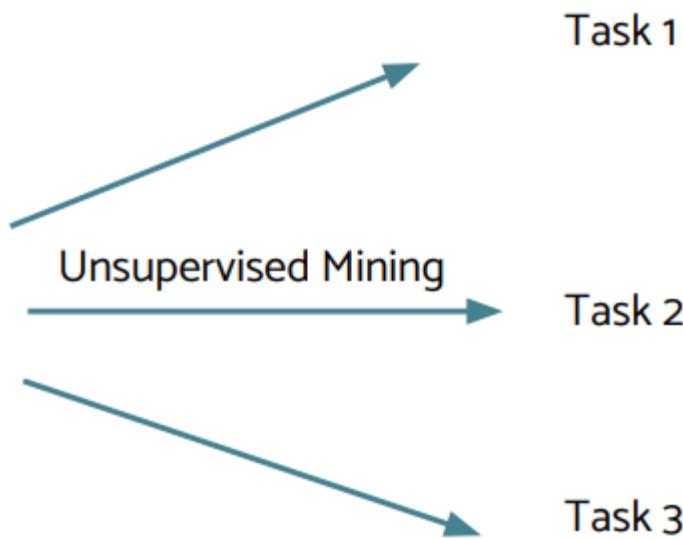
IndicGLUE (*Indic General Language Understanding Evaluation Benchmark*)

Task Type	Task	N	Languages
Classification	News Article Classification	10	bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Headline Classification	4	gu, ml, mr, ta
	Sentiment Analysis	2	hi, te
	Discourse Mode Classification	1	hi
Diagnostics	Winograd Natural Language Inference	3	gu, hi, mr
	Choice of Plausible Alternatives	3	gu, hi, mr
Semantic Similarity	Headline Prediction	11	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Wikipedia Section Titles	11	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Cloze-style Question Answering	11	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te
	Paraphrase Detection	4	hi, ml, pa, ta
Sequence Labelling	Named Entity Recognition	11	as, bn, gu, hi, kn, ml, mr, or, pa, ta, te
Cross-lingual	Cross-Lingual Sentence Retrieval	8	bn, gu, hi, ml, mr, or, ta, te

Creation of IndicGlue

Lack of Evaluation Datasets !

Data Sources
News Crawls
Wikipedia
Public Datasets



IndicGlue Tasks

6 Tasks
4 Types

Semantic

News Articles Headline Prediction
Wikipedia Section Title Prediction
Article Genre Classification

[News Crawls](#)
[Wikipedia](#)
[News Crawls](#)

Knowledge

Cloze-style multiple-choice QA

[Wikipedia](#)

Syntax

Named Entity Recognition

[Public Dataset](#)

Cross-lingual

Cross-Lingual Sentence Retrieval

[Public Dataset](#)

Additional Tasks (Paraphrase Detection, Movie Reviews etc.)

News headline prediction

Created From: News Crawls

IPL 2021: Australian Cricketers, Support Staff Expected To Head To Maldives

-ve

With their country shut for all those flying from India, the now-suspended IPL's Australian contingent, comprising players, support staff and commentators, is expected to head to Maldives before taking a connecting flight for home. The IPL was "indefinitely suspended" on Tuesday after multiple cases of COVID-19 emerged from Kolkata Knight Riders, Delhi Capitals, SunRisers Hyderabad and Chennai Super Kings. There are 14 Australian players along with coaches and commentators who might now take a detour as the Australian government has imposed strict sanctions for people returning from India.

Careful Negative Sampling

SRH vs MI, IPL 2021: SunRisers Hyderabad Players To Watch Out For

-ve

Bottom-placed SunRisers Hyderabad take on a high-flying Mumbai Indians team at the Arun Jaitley Stadium in Delhi on Tuesday. SunRisers Hyderabad have had a torrid time in IPL 2021 so far, winning a solitary game after playing seven matches. They have just two

Task: Predict the correct headline

IPL 2021: Mayank Agarwal's 99* In Vain As Delhi Capitals Thrash Punjab Kings To Go Top Of The Table

+ve

Shikhar Dhawan's delightful 69 dwarfed Mayank Agarwal's unbeaten 99 as Delhi Capitals defeated Punjab Kings by seven wickets in the IPL, on Sunday to go atop the points table. Agarwal, leading the side in the absence of regular skipper K L Rahul, used the straight bat effectively in his lone hand to take Punjab Kings to 166 for six. Delhi Capitals hardly broke a sweat in the run chase, cantering to victory in 17.4 overs overs, their sixth win in eight matches.

Input

Sri Lanka All-Rounder Thisara Perera Bids Adieu To International Cricket

-ve

Sri Lankan all-rounder Thisara Perera, on Monday, announced his retirement from international cricket with immediate effect. In a letter to Sri Lanka Cricket (SLC), Perera said that he wanted to focus on his family, before adding that it was the right time for him

Cloze-style multiple choice question

Created From: Wikipedia

Task: Predict the masked entity

Homi Bhabha was born in 1949 in Mumbai to a Parsi family. After receiving his early education at St. Mary's, he went on to graduate from Bombay University . He then moved to [MASK] for higher education . He received his MA and M.Phil degrees from Oxford University .

Candidate 1: Britain [correct answer]

Candidate 2: India

Candidate 3: Chicago

Candidate 4: Pakistan

Article Genre Classification

Created From: News Crawl

Task: Predict the genre of news article

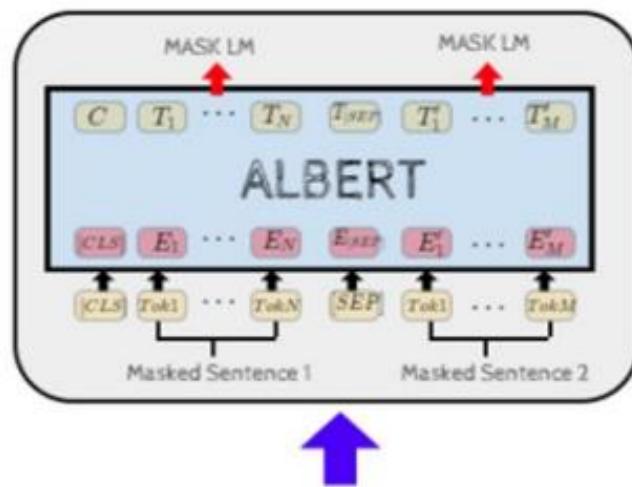
IPL 2021: Mayank Agarwal's 99* In Vain As Delhi Capitals Thrash Punjab Kings To Go Top Of The Table

Shikhar Dhawan's delightful 69 dwarfed Mayank Agarwal's unbeaten 99 as Delhi Capitals defeated Punjab Kings by seven wickets in the IPL, on Sunday to go atop the points table. Agarwal, leading the side in the absence of regular skipper K L Rahul, used the straight bat effectively in his lone hand to take Punjab Kings to 166 for six. Delhi Capitals hardly broke a sweat in the run chase, cantering to victory in 17.4 overs overs, their sixth win in eight matches.

Category: Sports

=> Mined from URL

IndicBERT



ପ୍ରଥିବା ଓ ଅ
ମୁମର୍ଚ ନେତ୍ର

Joint Pre-training

[ai4bharat/indic-bert · Hugging Face](#)

- Pre-trained Indic LM for **NLU applications**
- Large Indian language content (8B tokens)
 - 11 Indian languages
 - + Indian English content
- Multilingual Model
- Compact Model (~20m params)
- Competitive/better than mBERT/XLM-R
- Simplify fine-tune for your application
- 10k downloads per month on HuggingFace

Natural Language Generation



Machine
Translation



Automatic
Summarization

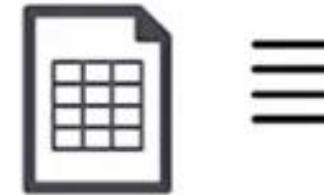
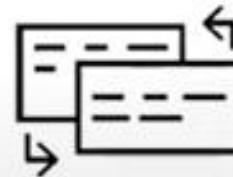


Table-to-Text
Generation



Dialog
Generation



Paraphrase
Generation

...

Missing in Indic Languages



Pretraining
Data and
Model

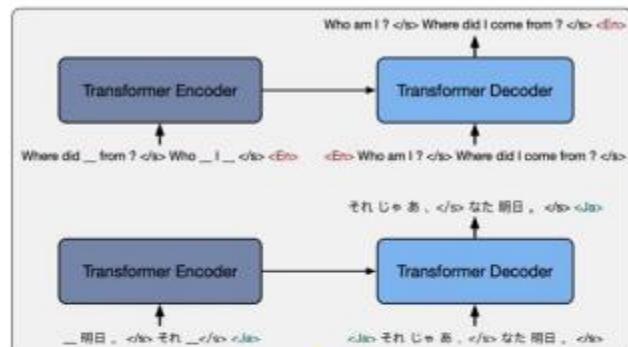


NLP Training
Data



Models across
IN-22

IndicBART



ପ୍ରି ହି ବା ଓ ଅ
ଗୁ ମ ଟ ତେ ଦିତ
Joint Pre-training

- Pre-trained Indic S2S for **NLG applications**
- Large Indian language content (8B tokens)
 - 11 Indian languages
 - + Indian English content
- Multilingual Model
- Compact Model (~224m params)
- **Single Script**
- Competitive with mBART50 for MT and summarization
- Simply **fine-tune** for your application

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, Pratyush Kumar. IndicBART: A Pre-trained Model for Natural Language Generation of Indic Languages. Arxiv preprint 2109.02903. 2021.

Train Indic BART

1 Leverage IndicCorp with data in 11 langs to train IndicBART

2 Exploit lang. similarity by script unification

3 Devise methods to auto-create NLG training data

Train IndicBART on IndicCorp



450M input
sentences of
training data



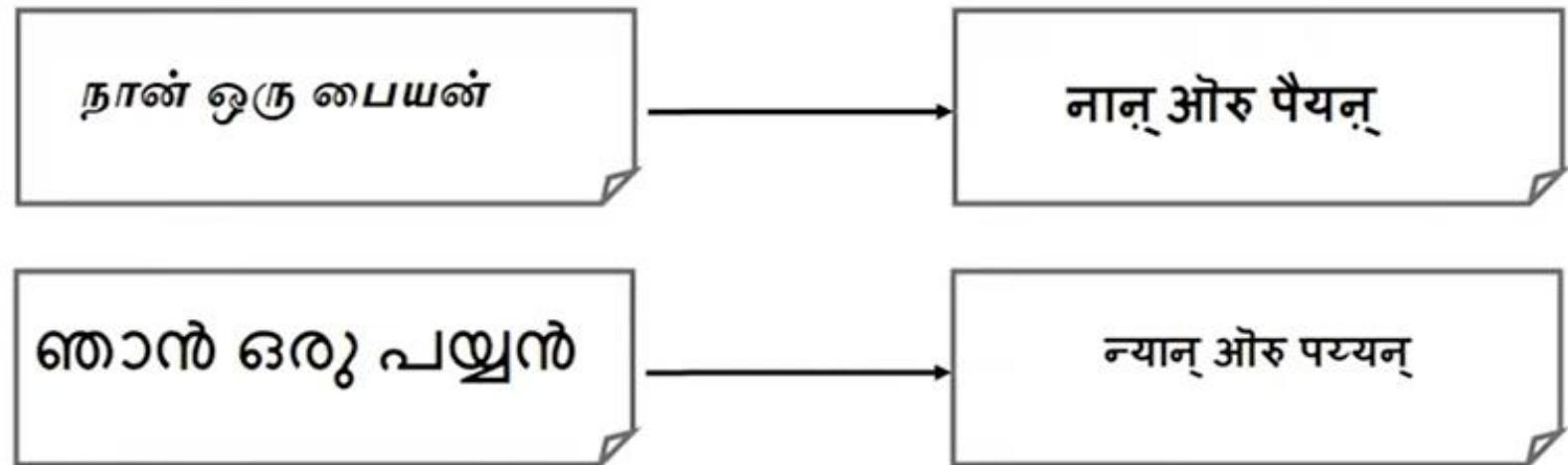
Compact
models with
244M
params



Covers 11
Indian
languages

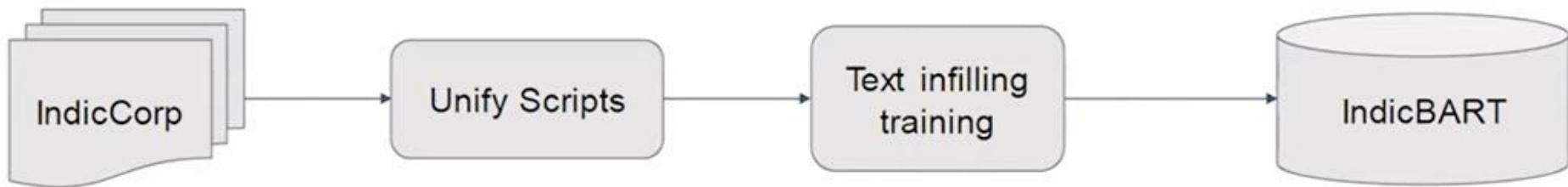
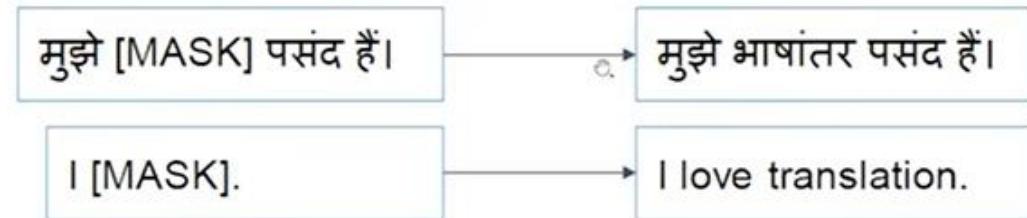
Script Unification

- Many languages need large vocabulary
- Script unification by converting to Devanagari
 - Increased vocabulary sharing
 - Compact vocabularies for compact models



IndicBART Training

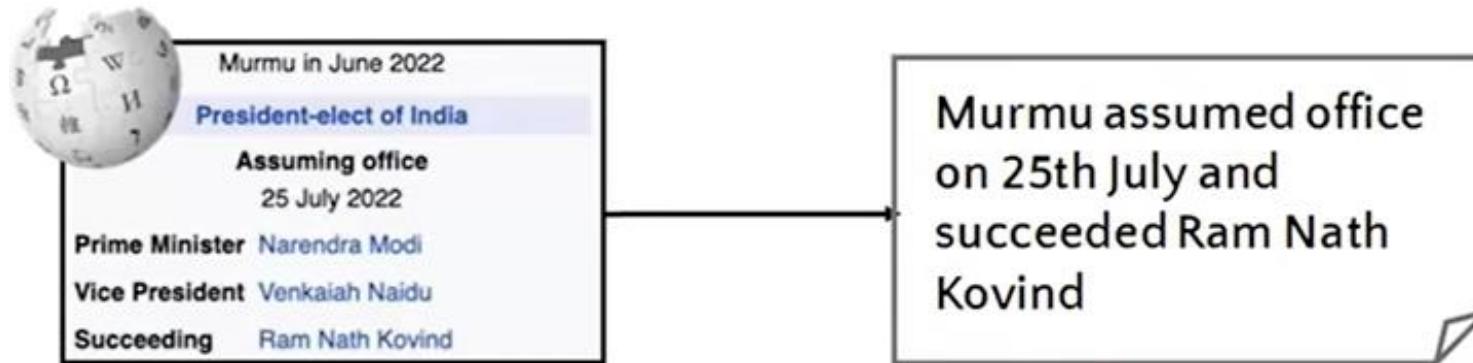
- Train models to do:
(text infilling)



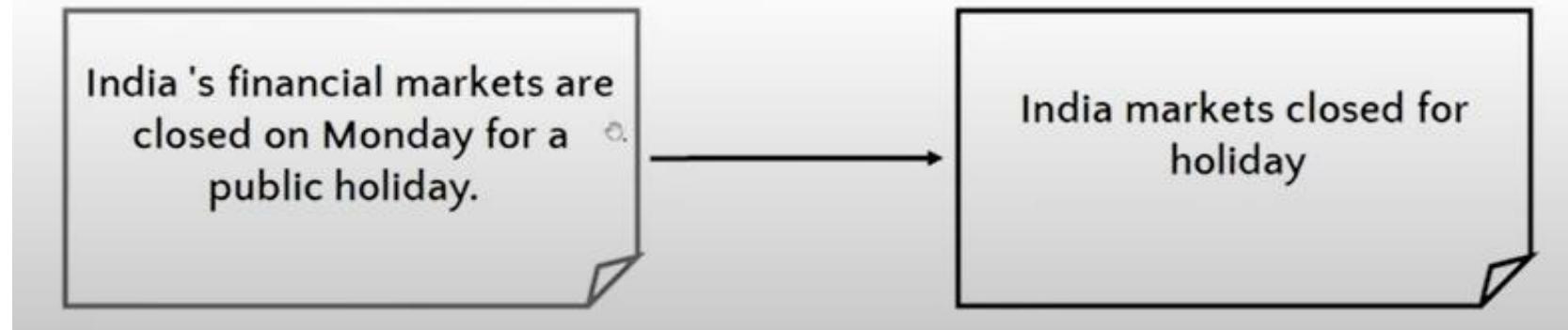
- IndicBART learns to infer a variation of input.
 - Learns generic NLG → Reduces need for task data (fine-tuning)
 - Variations: IndicALBART (compact)

Methods for creating training data

BIOGRAPHY GENERATION



SENTENCE SUMMARISATION



Methods for creating training data

PARAPHRASE GENERATION

दिल्ली विश्वविद्यालय, भारत में उच्च शिक्षा के लिए एक प्रतिष्ठित संस्थान है।



The University of Delhi is a prestigious institution for higher education in India.

Delhi University is one of the famous universities of the country.

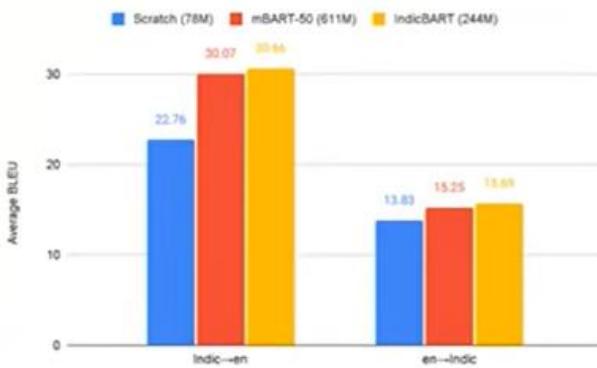
QUESTION GENERATION



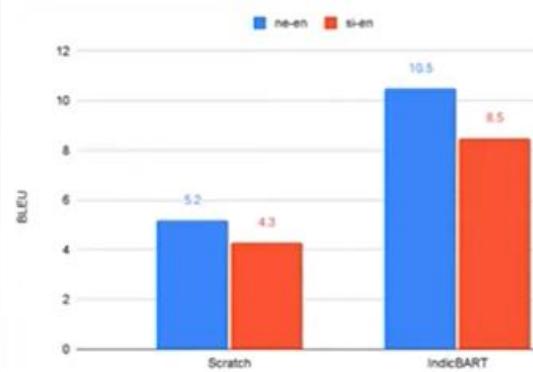
SQuAD

INDICTRANS

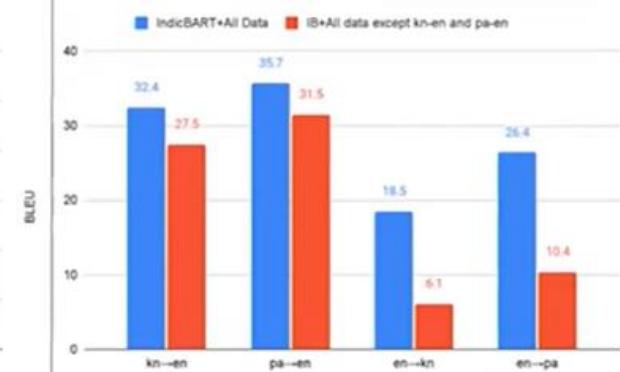
Machine Translation Bleu Score



- Large impact of pre-training
 - *Indic→En*: 22.76→30.66
 - *En→Indic*: 13.83→15.69
- Indic→En gains more than En→Indic

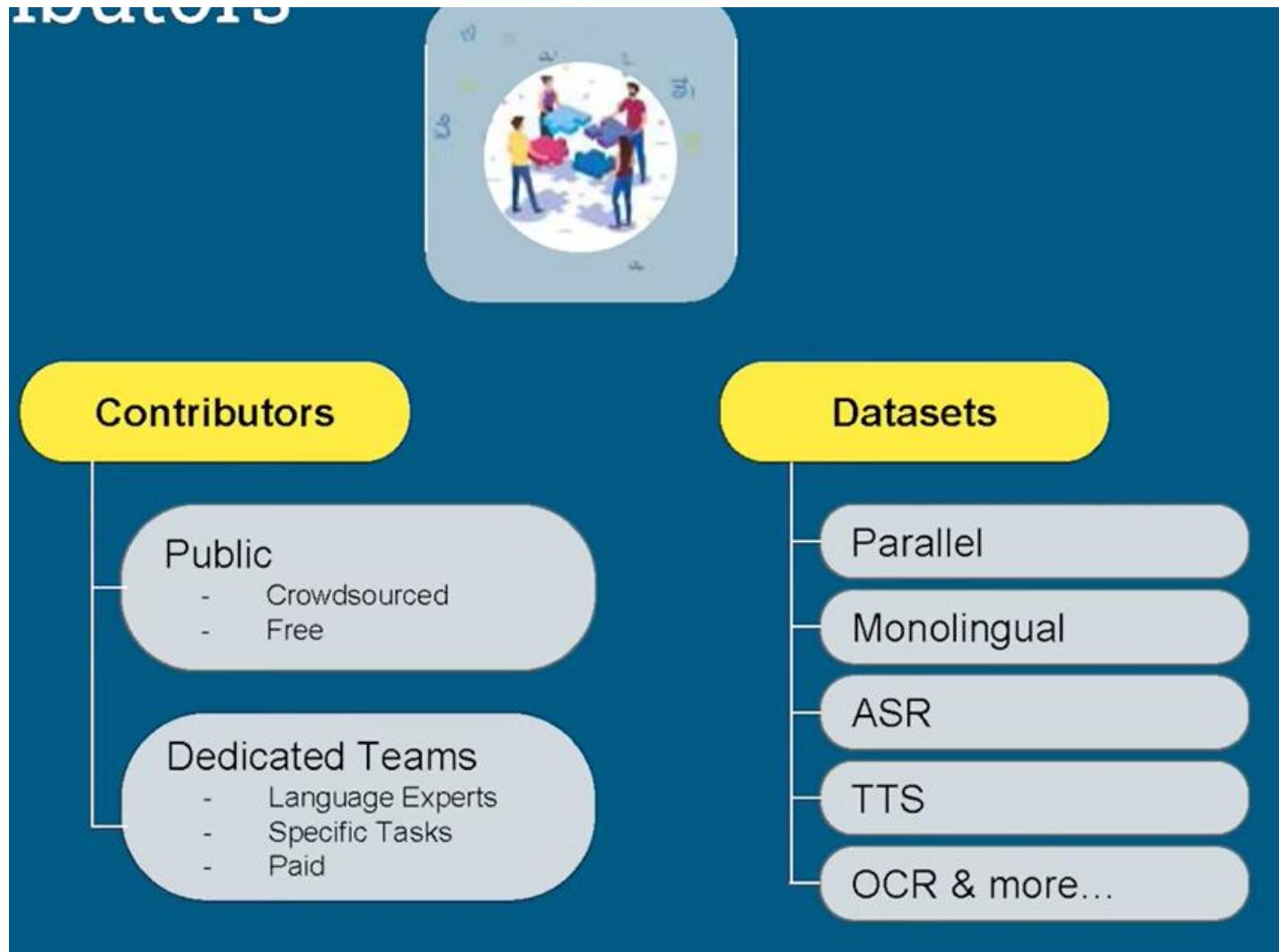


- IndicBART helps Nepali and Sinhala translation
- Both were unseen by IndicBART



- IndicBART helps unseen language translation
- Punjabi and Kannada data not used
 - Can still translate

Datasets



Bhasha daan

Help to build an open repository of data to digitally enrich your language



Enrich your language by typing the audio that you hear



Contribute by typing what you hear on audio

ASR Datasets



Enrich your language by donating your voice



Contribute your voice by recording the sentence

TTS Datasets



Enrich your language by translating text

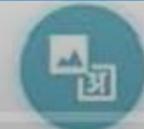


Contribute by translating the prompted text in your language

Parallel Datasets



Enrich your language by typing the text you see



OCR Datasets

Summary

- Deep Learning presents a unique opportunity to build NLP technologies at scale for Indian languages
 - Utilizing language relatedness is important to this mission
 - The orthographic similarity of Indian languages is a strong starting point for utilizing language relatedness.
 - Contact as well as genetic relatedness are useful in the context of Indian languages.
 - Multilingual pre-trained models trained on large corpora needed for transfer learning in NLU and NLG tasks
-

References

- https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00452/109468/Samanantar-The-Largest-Publicly-Available-Parallel
- <https://www.youtube.com/watch?v=QwYPOd1eBtQ&t=387s>
- <https://www.youtube.com/watch?v=I3TCBVBdqLo>
- https://www.youtube.com/watch?v=6Z69VW5clfk&list=PLKH1vOqy5KEsttDS4AftVcvHhlFw_awXa
- Indic Transliteration
- <https://www.youtube.com/watch?v=I3TCBVBdqLo>
- Multilingual Neural MT
- <https://www.youtube.com/watch?v=BdZeN-6TYzs>
- <https://www.youtube.com/@ai4bharat/featured>

<https://bhashini.gov.in/ulca/search->

<model/656f0d0700d64169e2f8f3f8/model>

<https://towardsdatascience.com/english-to-hindi->

[neural-machine-translation-7cb3a426491f](#)

[googleTranslationIndianLanguages -](#)

[Colaboratory](#)

Indic NER- Dataset

- Naamapadam Dataset
 - Large-Scale NER dataset for 11 Indic languages
 - As, Bn, Gu, Hi, Kn, Mi, Mr, Or, Pa, Ta, Te
 - Automated Creation via entity projection
 - Human annotated test-set for 8 Indic languages
 - Bn, Hi, Kn, Mi, Mr (large)
 - Ta, Te, Gu (small)
 - Multilingual IndicNER model
 - 11 Indic languages (As, Bn, Gu, Hi, Kn, Mi, Mr, Or, Pa, Ta, Te)
 - Compact 159.05 M parameters
 - Publicly available models and code
-

Named Entity Recognition

The task of identifying and extracting named entities in a given piece of text

For example,

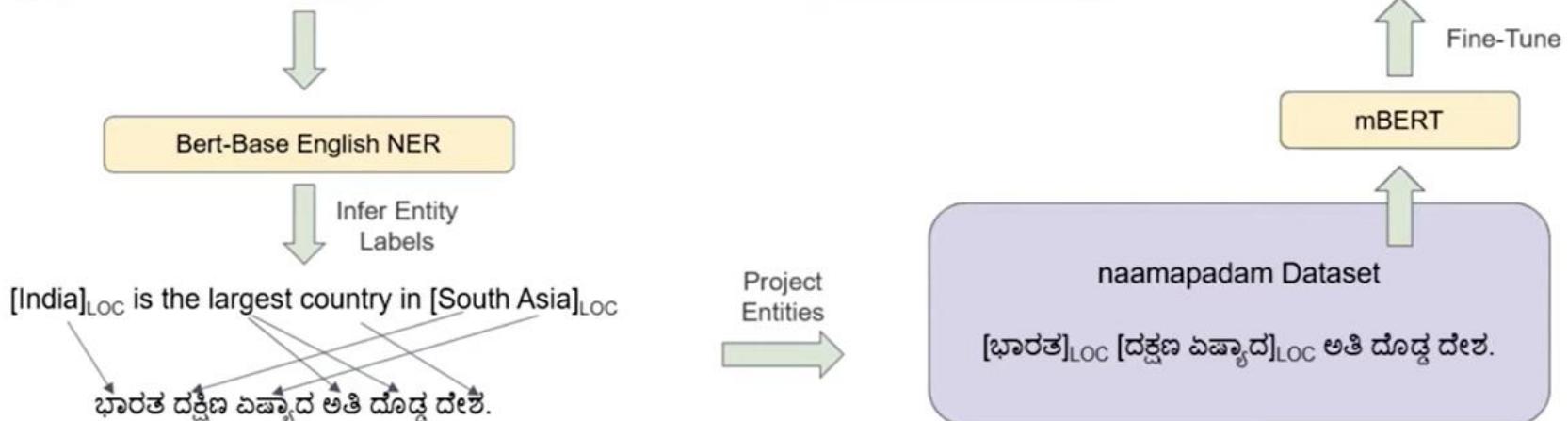
[Nilekani Center]_{LOCATION} at [AI4Bharat]_{ORGANIZATION} will be launched on [28th July]_{DATE} at [IIT Madras]_{ORGANIZATION}

Challenges in Indic languages:

- Lack of capitalization feature
- Ambiguity between Proper nouns and common nouns
- Morphological variations
- Small labelled data

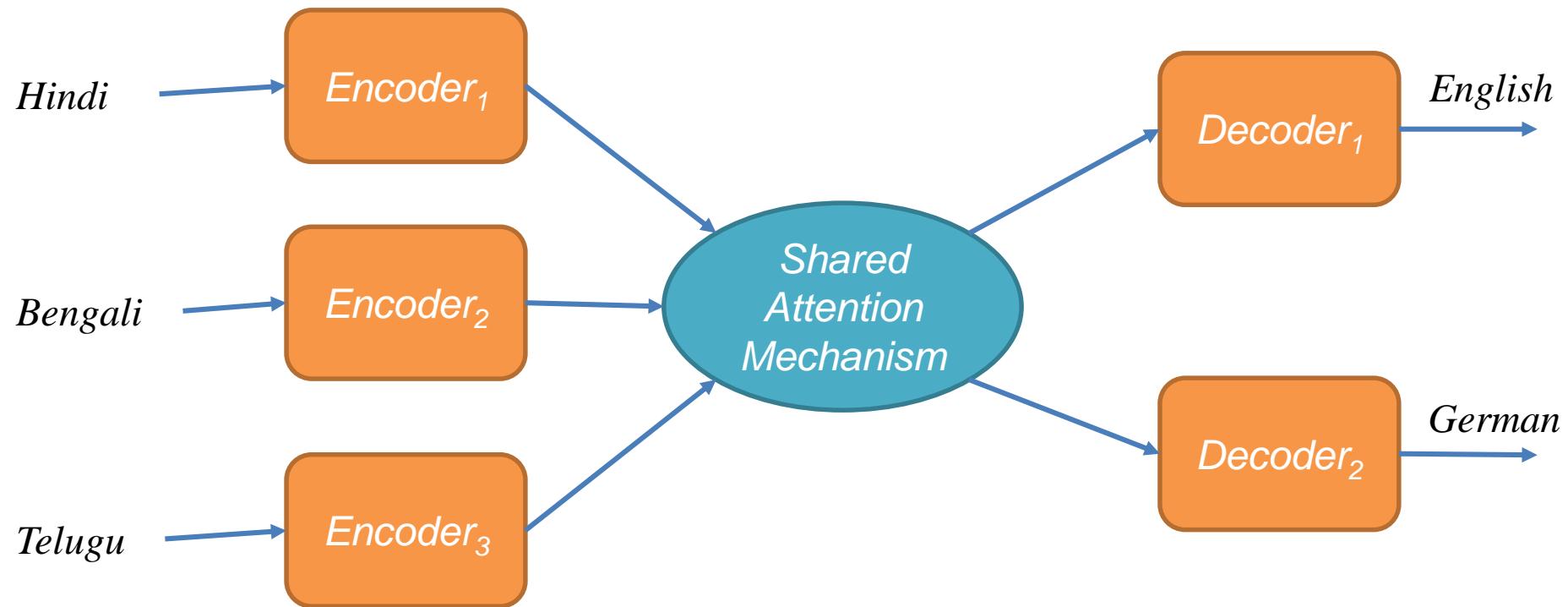
Indic NER Model

Samanantar Parallel Corpus	
English	Indic Language
India is the largest country in South Asia.	भारत दक्षिण एशिया का सबसे बड़ा देश है।
Mithali Raj was the captain of Indian women's cricket.	मिताली राज भारतीय महिला क्रिकेट की कप्तान हैं।
...	...



Multilingual Neural Translation

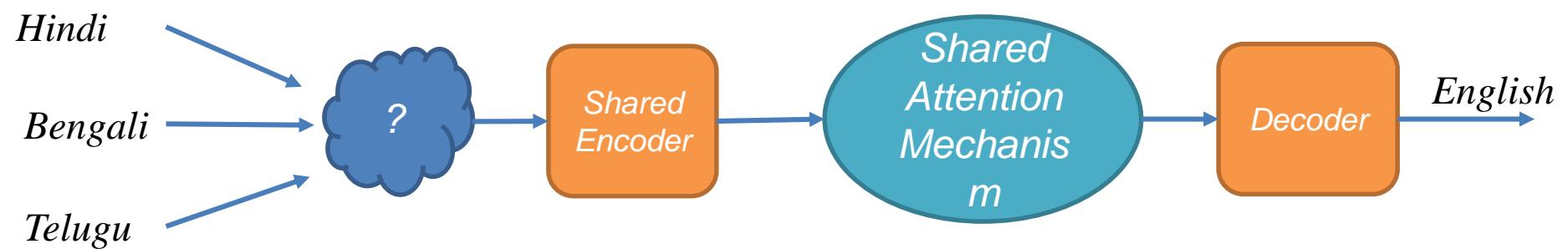
(Firat et al., 2016; Johnson et al., 2017)



Translate unseen language pairs → Zeroshot Translation

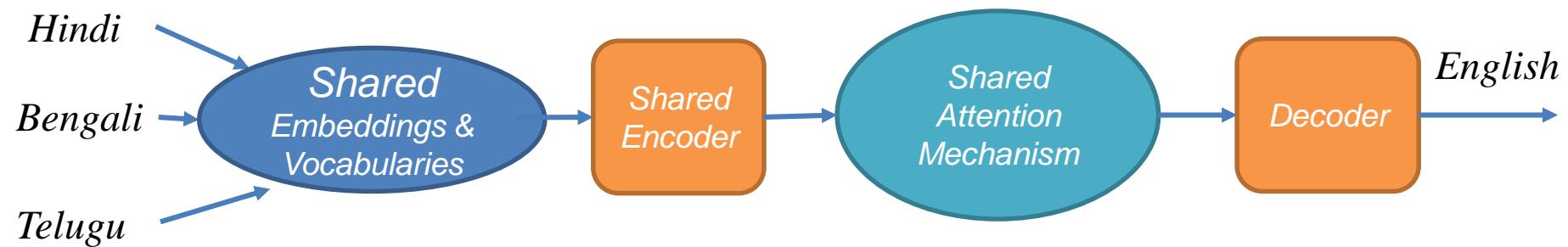
Shared Encoder

(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017)



Shared Encoder

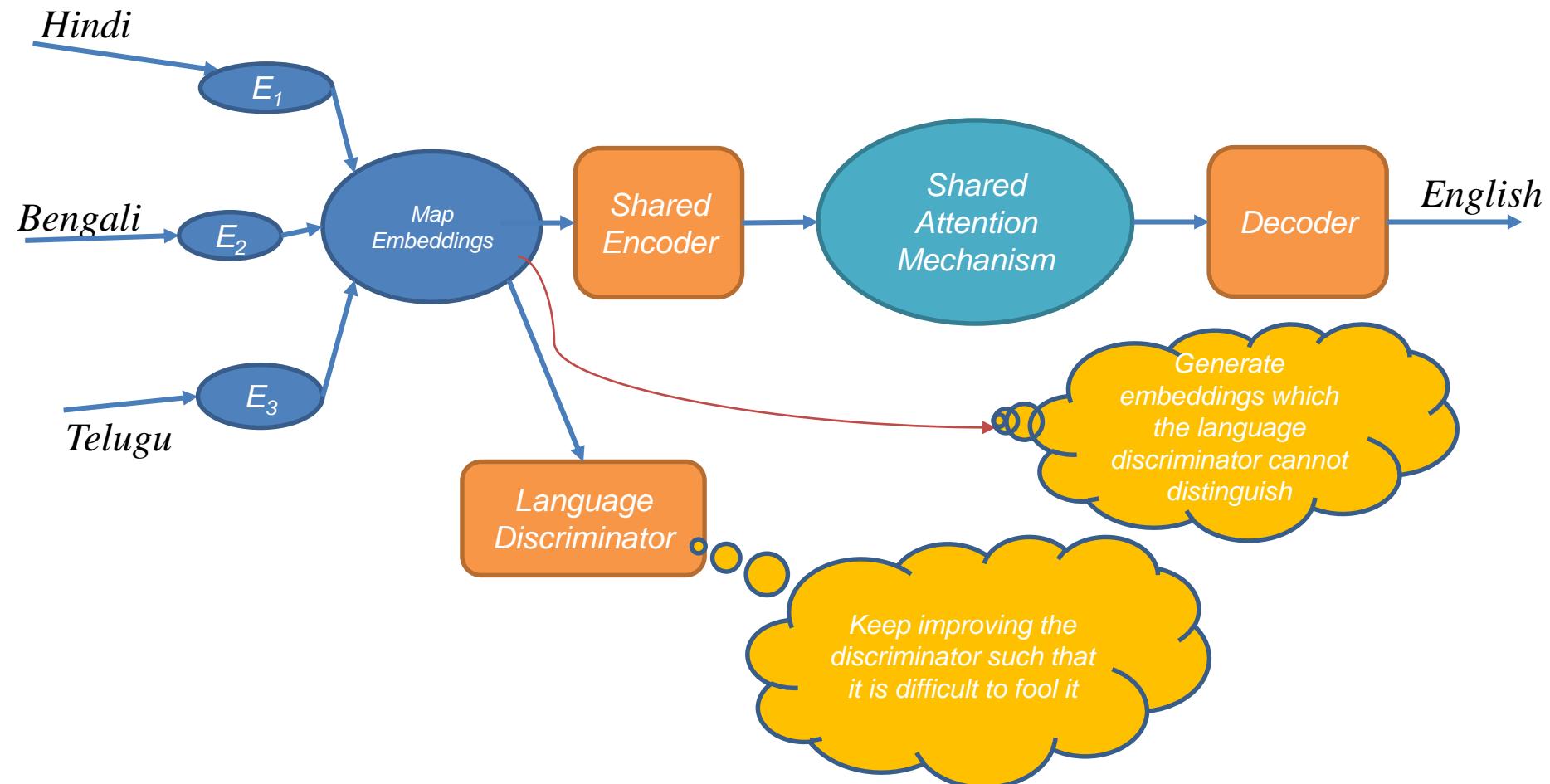
(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017)



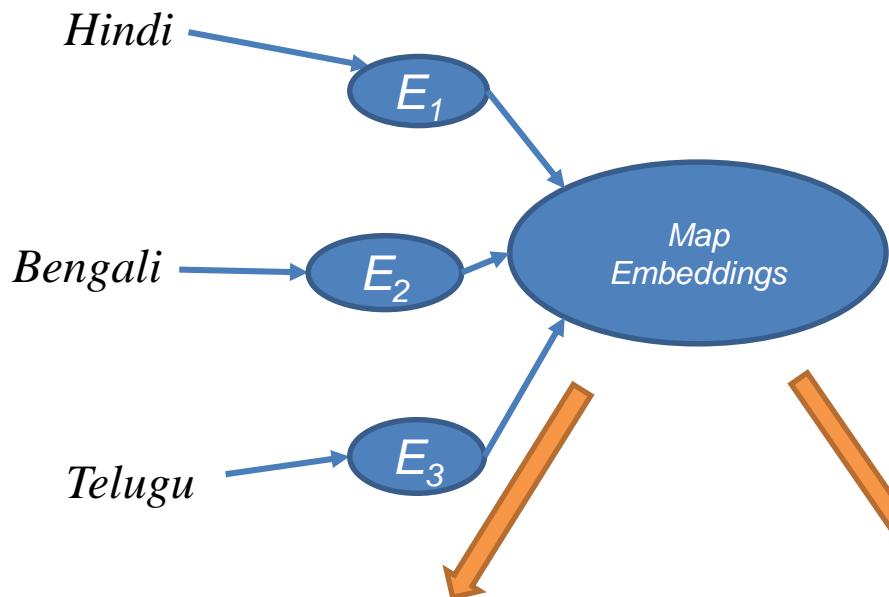
Shared Encoder with Adversarial Training



(Joty et al., 2017)



Learning Multilingual mappings/embeddings



Offline Mapping of embeddings

$$e_2 = A_{21}e_1$$

Joint training for multilingual embeddings

$$A_1 e_1 = A_2 e_2$$

The key to multilingual learning

- *Text Classification (sentiment analysis, Question matching)*
- *Sequence Tagging (POS, NER, etc.)*
- *Sequence to Sequence Learning (Machine Translation, Transliteration, etc)*



Needs parallel corpora or bilingual dictionaries



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 12: Information Extraction-Named Entity Recognition

Date – 3rd March 2024

These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philip Koehn**, Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

Agenda

- Information Extraction
- Typical IE Pipeline
- Named Entity Recognition
- Challenges in NER
- NER Approaches
- MEMM
- CRF
- Neural NER
- Evaluation of NER

Information Extraction (IE)

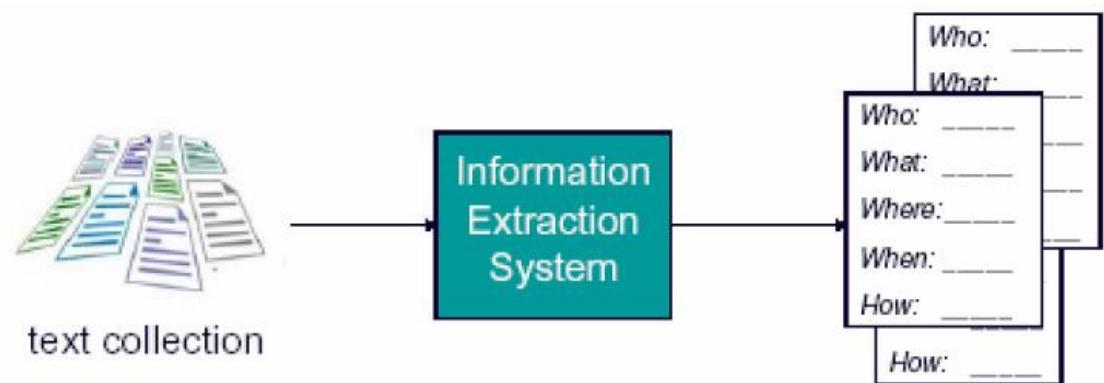
- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
 - Newspaper articles
 - Web pages
 - Scientific articles
 - Newsgroup messages
 - Classified ads
 - Medical notes

Information Extraction vs. NLP?

- Information extraction is attempting to find *some* of the structure and meaning in the hopefully template driven web pages.
- As IE becomes more ambitious and text becomes more free form, then ultimately we have IE becoming equal to NLP.
- Web does give one particular boost to NLP
 - *Massive corpora..*

Information Extraction (IE)

- Identify specific pieces of information (data) in an unstructured or semi-structured text
- Transform unstructured information in a corpus of texts or web pages into a structured database (or templates)
- Applied to various types of text, e.g.
 - Newspaper articles
 - Scientific articles
 - Web pages



Source: J. Choi, CSE842, MSU

A Typical IE Processing Pipeline

*Named Entity Recognition (NER) &
Shallow Parsing*



Reference Resolution



Relation Detection & Classification



Event Detection & Classification



Template Filling

What is Information Extraction

As a task:

Filling slots in a database from sub-segments of text.

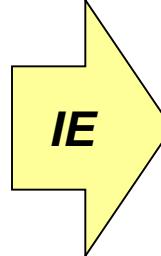
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is Information Extraction?

**As a family
of techniques:**

**Information Extraction =
segmentation + classification + association**

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Adapted from slide by William Cohen

Microsoft Corporation

CEO

Bill Gates

Microsoft

Gates

Microsoft

Bill Veghte

Microsoft

VP

Richard Stallman

founder

Free Software Foundation

aka "named entity extraction"

What is Information Extraction



*A family
of techniques:*

**Information Extraction =
segmentation + classification + association**

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Adapted from slide by William Cohen

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is Information Extraction

*A family
of techniques:*

***Information Extraction =
segmentation + classification + association***

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Adapted from slide by William Cohen

Microsoft Corporation

CEO

Bill Gates

Microsoft

Gates

Microsoft

Bill Veghte

Microsoft

VP

Richard Stallman

founder

Free Software Foundation

What is Information Extraction

*A family
of techniques:*

***Information Extraction =
segmentation + classification + association***

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Adapted from slide by William Cohen

Microsoft Corporation

CEO

Bill Gates

Microsoft

Gates

Microsoft

Bill Veghte

Microsoft

VP

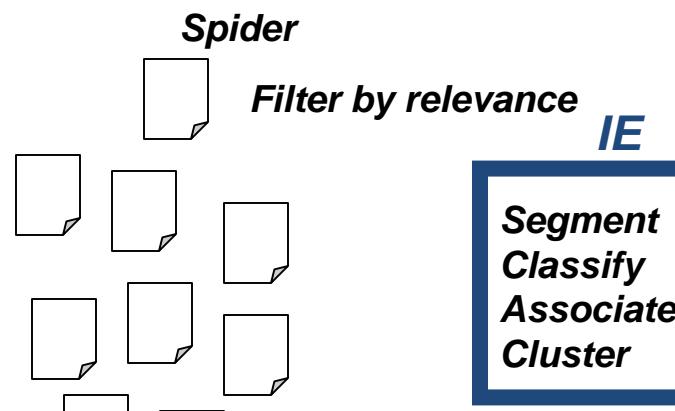
Richard Stallman

founder

Free Software Foundation

IE in Context

Create ontology

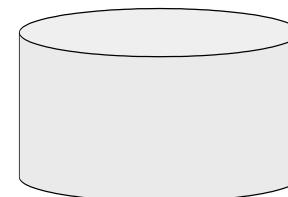


IE

**Segment
Classify
Associate
Cluster**

Train extraction models

Load DB



Label training data

*Query,
Search*

Data mine

Landscape of IE Tasks: Degree of Formatting



Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Non-grammatical snippets, rich formatting & links

Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor.			
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			
Brock, Oliver	(413) 577-0334	oli@cs.umass.edu	CS246
Assistant Professor.			
Clarke, Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS304
Professor.			
Software verification, testing, and analysis; software architecture and design.			
Cohen, Paul R.	(413) 545-3638	cohen@cs.umass.edu	CS278
Professor.			
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.			

Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press
- Contact**
- General information
- Directions maps

Tables

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty Joseph Y. Halpern, Cornell University				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis	Neural Networks	Games
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Solving through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van Nuffelen</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, Thomas Eiter, and Georg Gottlob</i>	179: Knowledge Extraction and Comparison from Local Function Networks <i>Kenneth McGarry, Stefan Wermter, and John MacIntyre</i>	71: Iterative Widening <i>Tristan Cazenave</i>
549: Online-Execution of ccGolog Plans <i>Henrik Grosskreutz</i>	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

Landscape of IE Tasks: Intended Breadth of Coverage



Web site specific

Formatting

Amazon.com Book Pages

The image shows two side-by-side screenshots of Amazon.com product pages. The left page is for 'Machine Learning' by Tom M. Mitchell, showing a 'Look Inside!' button and a 'Great Buy' badge. The right page is for 'Learning in Graphical Models' by Michael Irwin Jordan, also featuring a 'Look Inside!' button and a 'Great Buy' badge. Both pages include standard Amazon navigation like 'SEARCH', 'BROWSE SUBJECTS', and 'VIEW CART'.

Genre specific

Layout

Resumes

Two resumes are shown side-by-side. The top resume is for Jason D. M. Rennie, a researcher at the Massachusetts Institute of Technology. It includes sections for 'Research Interests' (mentioning automated analysis of data for classification, estimation, and acquiring new knowledge) and 'Education' (MIT). The bottom resume is for L. Douglas Baker, a researcher at Carnegie Mellon University. It includes sections for 'Home Address', 'Office Address', 'Office Phone', 'Home Page', 'Objective', 'Education' (Carnegie Mellon University), and 'Research Experience' (Carnegie Mellon University).

Wide, non-specific

Language

University Names

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty Joseph Y. Halpern, Cornell University				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis	Neural Networks	Games
739: A Logical Account of Causal and Topological Maps Emilio Remolina and Benjamin Kuipers	116: A-System: Solving through Abduction Marc Denecker, Antonis Kakas, and Bert Van	758: Title Generation for Machine-Translated Documents Rong Jin and Alexander G. Hauptmann	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories Kenneth McGarry, Stefan Marco Cadoli, and Bert Van	179: Knowledge Extraction and Comparison from Local Function Networks Tristan Cazenave	71: Iterative Widening Tristan Cazenave

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

- Press Contact
- General information
- Directions maps

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

Adapted from slide by William Cohen

Information Extraction

- 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

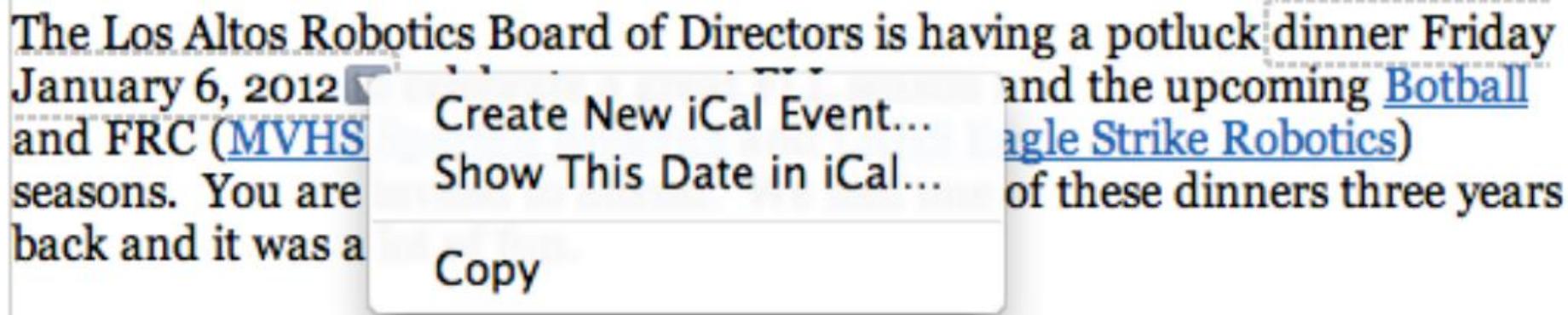
Information Extraction



- 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

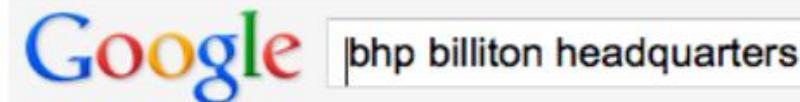
Low level Information Extraction

- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing



- Often seems to be based on regular expressions and name lists

Low level Information Extraction



Search

About 123,000 results (0.23 seconds)

Everything

Best guess for BHP Billiton Ltd. Headquarters is **Melbourne, London**

Images

Mentioned on at least 9 websites including [wikipedia.org](#), [bhpbilliton.com](#) and [bhpbilliton.com - Feedback](#)

Maps

[**BHP Billiton - Wikipedia, the free encyclopedia**](#)

Videos

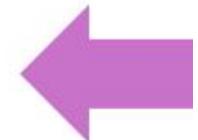
[en.wikipedia.org/wiki/BHP_Billiton](#)

News

Merger of BHP & Billiton 2001 (creation of a DLC). **Headquarters, Melbourne, Australia (BHP Billiton Limited and BHP Billiton Group) London, United Kingdom ...**

Shopping

[History - Corporate affairs - Operations - Accidents](#)



Named Entity Recognition

- A very important sub-task: **find** and **classify** names in text, for example:
 - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition

- A very important sub-task: **find** and **classify** names in text, for example:
 - The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: **find** and **classify** names in text, for example:
 - The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organization

Evolution of NER



Traditional

- Rules
- Regular Expressions
- **Gazetteers**

Statistical

- Word-based models – PMI, log-likelihood.
- Sequence models – **Conditional Random Fields**

Neural

- Bi-LSTM
- **Bi-LSTM+CRF**
- Transformer based Models

Rule based NER

The textbook gives an example of an iterative approach that makes multiple passes over the text:

- Pass 1: Use high-precision rules to label (a small number of) unambiguous mentions
- Pass 2: Propagate the labels of the previously detected named entities to any mentions that are substrings (or acronyms?) of these entities
- Pass 3: Use application-specific name lists to identify further likely names (as features?)
- Pass 4: Now use a sequence labeling approach for NER, keeping the already labeled entities as high-precision anchors.

The basic ideas behind this approach (label propagation, using high-precision items as anchors) can be useful for other tasks as well.

NER Task

Task: Predict entities in a text

Foreign	ORG	
Ministry	ORG	
spokesman	O	Standard
Shen	PER	
Guofang	PER	}
told	O	evaluation
Reuters	ORG	is per entity,
:	:	<i>not</i> per token

Variations and Ambiguity in NE

- Variation of NEs.
 - Manmohan Singh, Manmohan, Dr. Manmohan Singh
- Ambiguity of NE types:
 - 1945 (date vs. time)
 - Washington (location vs. person)
 - May (person vs. month)
 - Tata (person vs. organization)



More complex problems in NER

Issues of style, structure, domain, genre etc.

- Punctuation, spelling, spacing, formatting,all have an impact

Dept. of Computing and Information Science

Manchester Metropolitan University

Manchester

United Kingdom

- > Tell me more about Leonardo
- > Da Vinci

Problems in NE Task Definition



- Category definitions are intuitively quite clear, but there are many grey areas.
- Many of these grey area are caused by **metonymy**.

Person vs. Artefact

Organisation vs. Location

Company vs. Artefact

Location vs. Organisation

NER Approaches

Statistical models:

- Maximum Entropy Markov Models (MEMMs)
- Conditional Random Fields (CRFs)

Neural models:

- Recurrent networks (or transformers) that predict a label at each time step, possibly with a CRF output layer.

ML Sequence model Approach

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

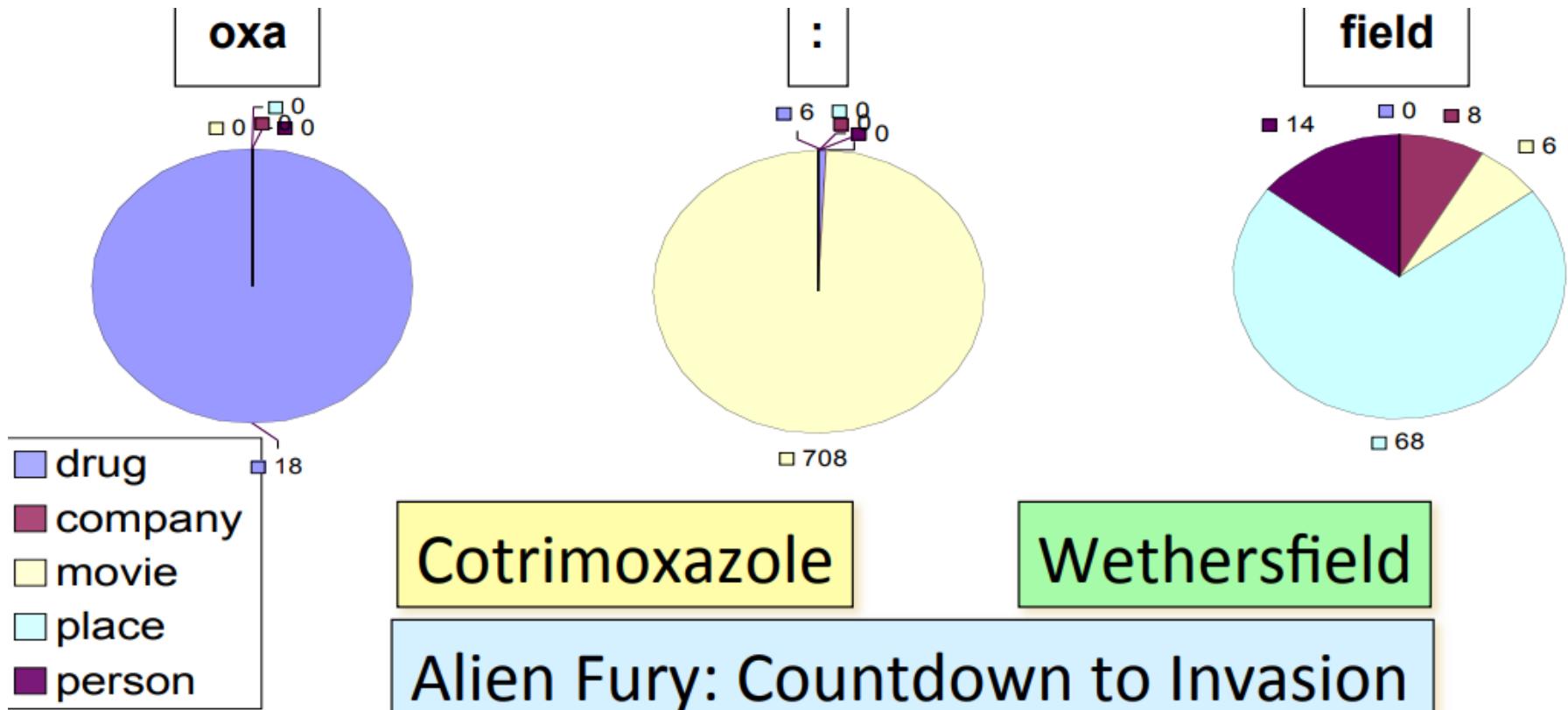
Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
's	O	O
new	O	O
painting	O	O

Features of sequence labeling

- Words
 - Current word (essentially like a learned dictionary)
 - Previous/next word (context)
- Other kinds of inferred linguistic classification
 - Part-of-speech tags
- Label context
 - Previous (and perhaps next) label

Features: Word Substrings



Features: Word Shapes

- Word Shapes
 - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

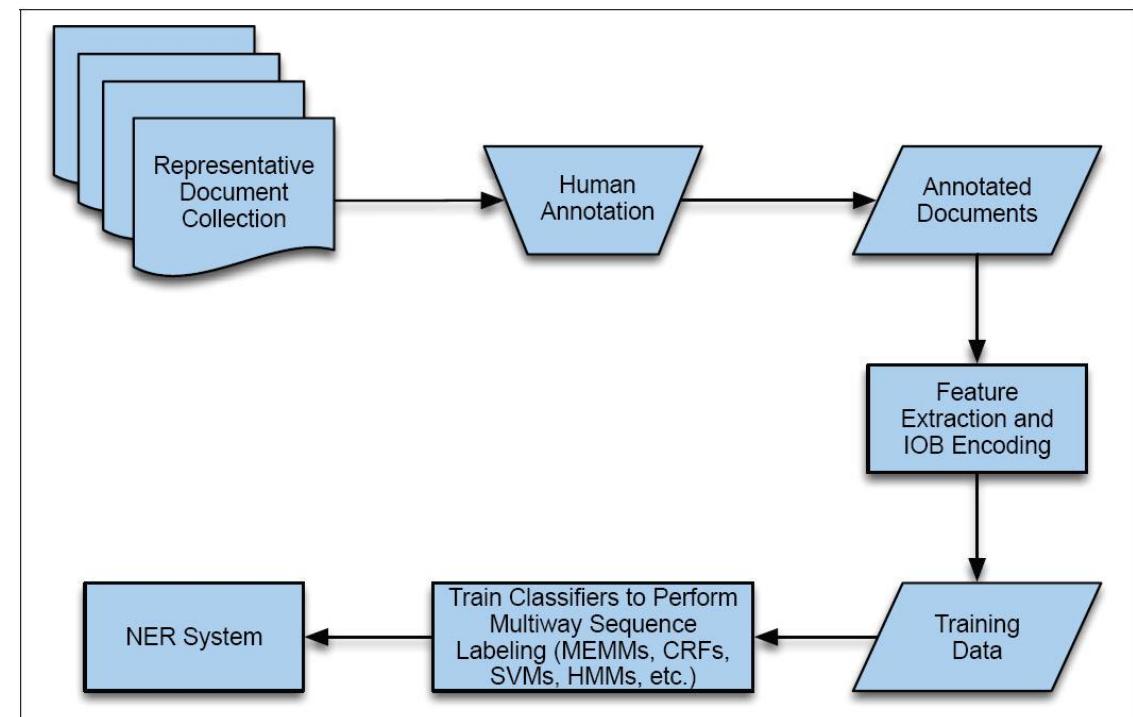
Named Entity Recognition

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad].

IOB notation

Word	POS	Chunk	EntityType
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

IE by statistical sequence labeling approach



Pierre Vinken , 61 years old , will join IBM 's board
as a nonexecutive director Nov. 29 .



[PERS Pierre Vinken] , 61 years old , will join
[ORG IBM] 's board as a nonexecutive director
[DATE Nov. 2] .

Task: identify all mentions of named entities
(people, organizations, locations, dates)

We define many new tags:

- **B-PERS, B-DATE, ...:** beginning of a mention of a person/date...
- **I-PERS, I-DATE, ...:** inside of a mention of a person/date...

```
[PERS Pierre Vinken] , 61 years old , will join  
[ORG IBM] 's board as a nonexecutive director  
[DATE Nov. 2] .
```



```
Pierre_B-PERS Vinken_I-PERS _O 61_O years_O old_O ,_O  
will_O join_O IBM_B-ORG 's_O board_O as_O a_O  
nonexecutive_O director_O Nov._B-DATE 29_I-DATE ._O
```

Biological NER

- There are a much wider range of entity types (semantic classes) in the biological domain

*[**tissue** Plasma] [**GP** BNP] concentrations were higher in both the [**population** *judo*] and [**population** marathon groups] than in [**population** controls], and positively correlated with [**ANAT** LV] mass as well as with deceleration time.*

Semantic class	Examples
Cell lines	<i>T98G, HeLa cell, Chinese hamster ovary cells, CHO cells</i>
Cell types	<i>primary T lymphocytes, natural killer cells, NK cells</i>
Chemicals	<i>citric acid, 1,2-diiodopentane, C</i>
Drugs	<i>cyclosporin A, CDDP</i>
Genes/proteins	<i>white, HSP60, protein kinase C, L23A</i>
Malignancies	<i>carcinoma, breast neoplasms</i>
Medical/clinical concepts	<i>amyotrophic lateral sclerosis</i>
Mouse strains	<i>LAFT, AKR</i>
Mutations	<i>C10T, Ala64 → Gly</i>
Populations	<i>judo group</i>

Biological NER (cont.)

- NER in this domain is particularly difficult because of the various forms which the names can take:
 - e.g. “insulin”, “ether a go-go”, “breast cancer associated 1”
 - Long names (thus multi-token boundary detection is needed)
 - Spelling/typographical variations
 - Abbreviations, symbols
 - (Of course) Ambiguity (common meaning or domain concepts)
- Extracted NEs are often mapped to **biomedical ontologies** (e.g. Gene Ontology, UMLS)

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

Q
A
Q
A
A
A
Q
A

**Text
segmen-
tation**

MEMM

- 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	%

Features	
w_0	22.6
w_{+1}	%
w_{-1}	fell
T_{-1}	VBD
$T_{-1}-T_{-2}$	NNP-VBD
hasDigit?	true
...	...

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

MEMM: POS Tagging

- Scoring individual labeling decisions is no more complex than standard classification decisions
 - We have some assumed labels to use for prior positions
 - We use features of those and the observed data (which can include current, previous, and next words) to predict the current label

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
...	...

MEMM: POS Tagging

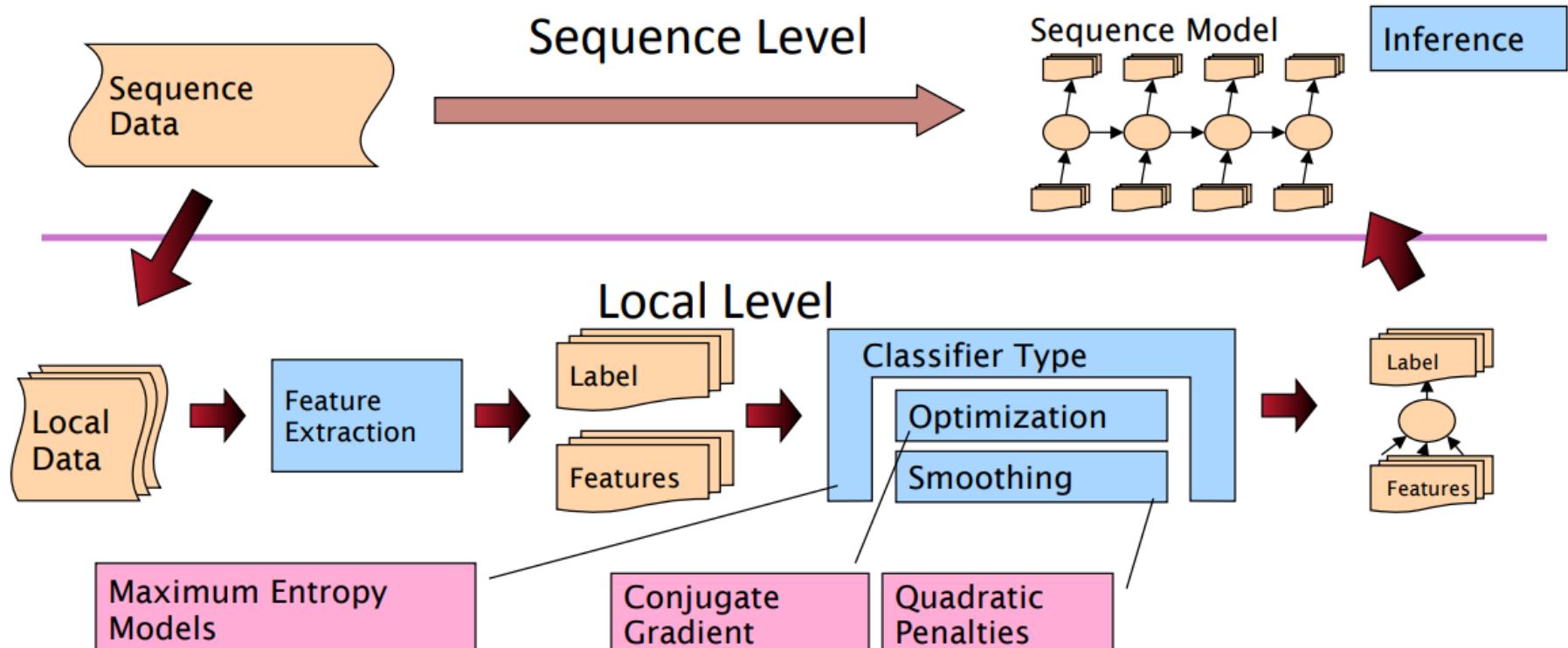
- POS tagging Features can include:
 - Current, previous, next words in isolation or together.
 - Previous one, two, three tags.
 - Word-internal features: word types, suffixes, dashes, etc.

Local Context					Decision Point
-3	-2	-1	0	+1	
DT	NNP	VBD	???	???	
The	Dow	fell	22.6	%	

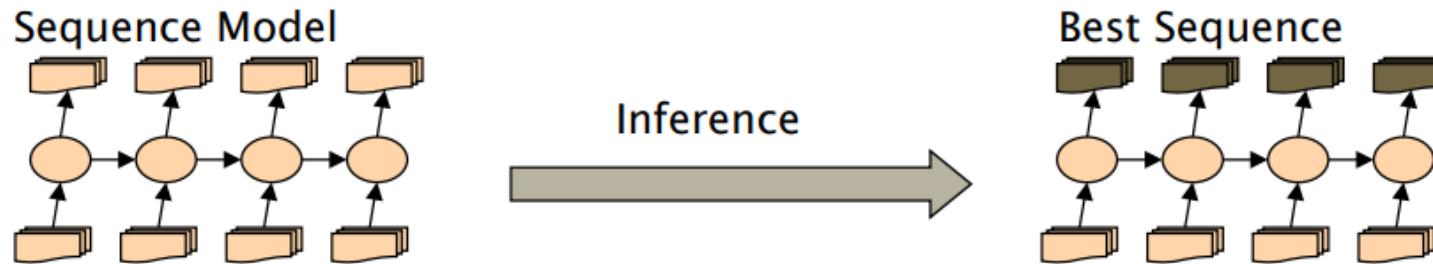
Features	
W ₀	22.6
W ₊₁	%
W ₋₁	fell
T ₋₁	VBD
T ₋₁ -T ₋₂	NNP-VBD
hasDigit?	true
...	...

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

MEMM: Inference



MEMM: Greedy Inference



Greedy inference:

- We just start at the left, and use our classifier at each position to assign a label
- The classifier can depend on previous labeling decisions as well as observed data

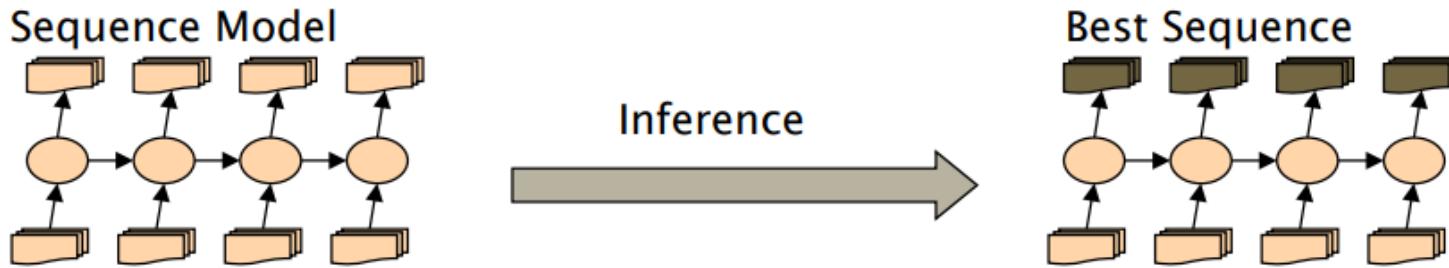
Advantages:

- Fast, no extra memory requirements
- Very easy to implement
- With rich features including observations to the right, it may perform quite well

Disadvantage:

- Greedy. We make commit errors we cannot recover from

MEMM: Beam Inference



Beam inference:

- At each position keep the top k complete sequences.
- Extend each sequence in each local way.
- The extensions compete for the k slots at the next position.

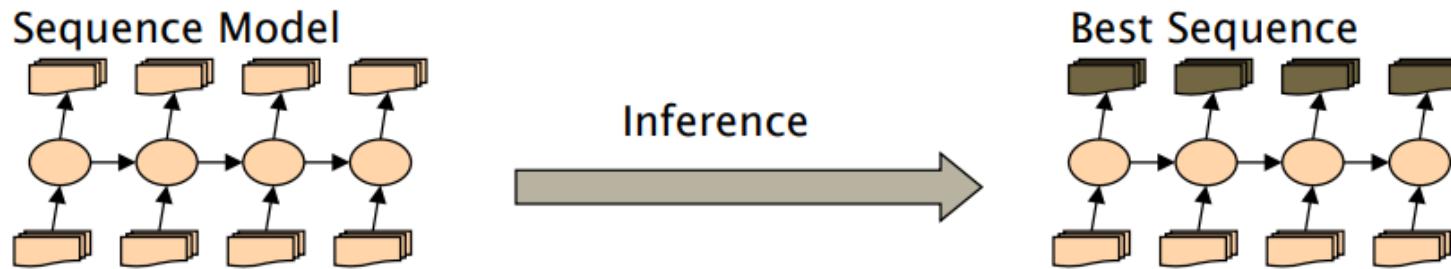
Advantages:

- Fast; beam sizes of 3-5 are almost as good as exact inference in many cases.
- Easy to implement (no dynamic programming required).

Disadvantage:

- Inexact: the globally best sequence can fall off the beam.

MEMM: Viterbi Inference



Viterbi inference:

- Dynamic programming or memoization.
- Requires small window of state influence (e.g., past two states are relevant).

Advantage:

- Exact: the global best sequence is returned.

Disadvantage:

- Harder to implement long-distance state-state interactions (but beam inference tends not to allow long-distance resurrection of sequences anyway).

Conditional Random Field (CRF)

Another sequence model: Conditional Random Fields (CRFs)

A whole-sequence conditional model rather than a chaining of local models.

$$P(c | d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

The space of c 's is now the space of sequences

- But if the features f_i remain local, the conditional sequence likelihood can be calculated exactly using dynamic programming

Training is slower, but CRFs avoid causal-competition biases

These (or a variant using a max margin criterion) are seen as the state-of-the-art these days ... but in practice usually work much the same as MEMMs.

Named Entity Types

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	Turing is a giant of computer science.
Organization	ORG	companies, sports teams	The IPCC warned about the cyclone.
Location	LOC	regions, mountains, seas	The Mt. Sanitas loop is in Sunshine Canyon.
Geo-Political	GPE	countries, states, provinces	Palo Alto is raising the fees for parking.
Entity			
Facility	FAC	bridges, buildings, airports	Consider the Golden Gate Bridge.
Vehicles	VEH	planes, trains, automobiles	It was a classic Ford Falcon.

Figure 18.1 A list of generic named entity types with the kinds of entities they refer to.

These types were developed for the news domain as part of NIST's Automatic Content Extraction (ACE) program.

Other domains (e.g. biomedical text) require different types (proteins, genes, diseases, etc.)

Feature based NER

identity of w_i , identity of neighboring words
embeddings for w_i , embeddings for neighboring words
part of speech of w_i , part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
 w_i contains a particular prefix (from all prefixes of length ≤ 4)
 w_i contains a particular suffix (from all suffixes of length ≤ 4)
 w_i is all upper case
word shape of w_i , word shape of neighboring words
short word shape of w_i , short word shape of neighboring words
presence of hyphen

Figure 18.5 Typical features for a feature-based NER system.

Train a sequence labeling model (MEMM or CRF),
using features such as the ones listed above for English

- Word Shape: replace all upper-case letters with one symbol (e.g. “X”), all lower-case letters with another symbol (“x”), all digits with another symbol (“d”), and leave punctuation marks as is (“L’Occitane → “X’Xxxxxxxxx”)
- Short Word Shape: remove adjacent letters that are identical in word shape
“L’Occitane → “X’Xxxxxxxxx” → “X’Xx””

Input Format – BIO Tagging

Barack Obama is 44th United States President.

Barack Obama	PER
United States	LOC

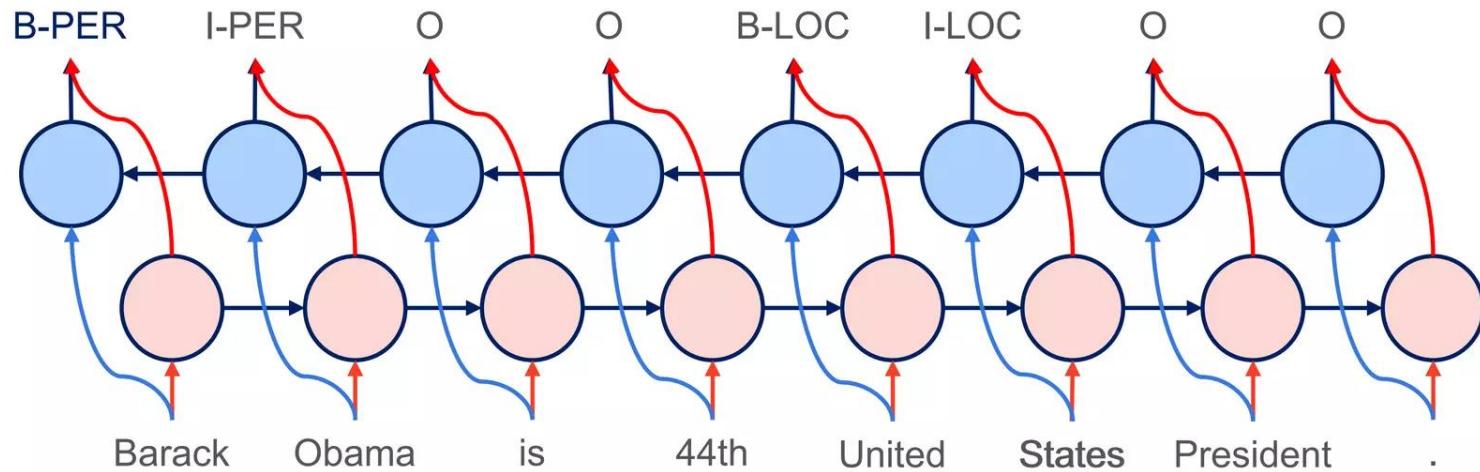
CoNLL

- **BIO** – Begin In Out.
 - Barack/**B-PER** Obama/**I-PER** is/O 44th/O United/**B-LOC** States/**I-LOC** President/O ./O
- **BILOU** – a tagging variant:
 - **U** – **Unit** token (for single token entities)
 - **L** – **Last** token in sequence, ex. Barack/B-PER Obama/L-PER

Barack	B-PER
Obama	I-PER
is	O
44 th	O
United	B-LOC
States	I-LOC
President	O
.	O



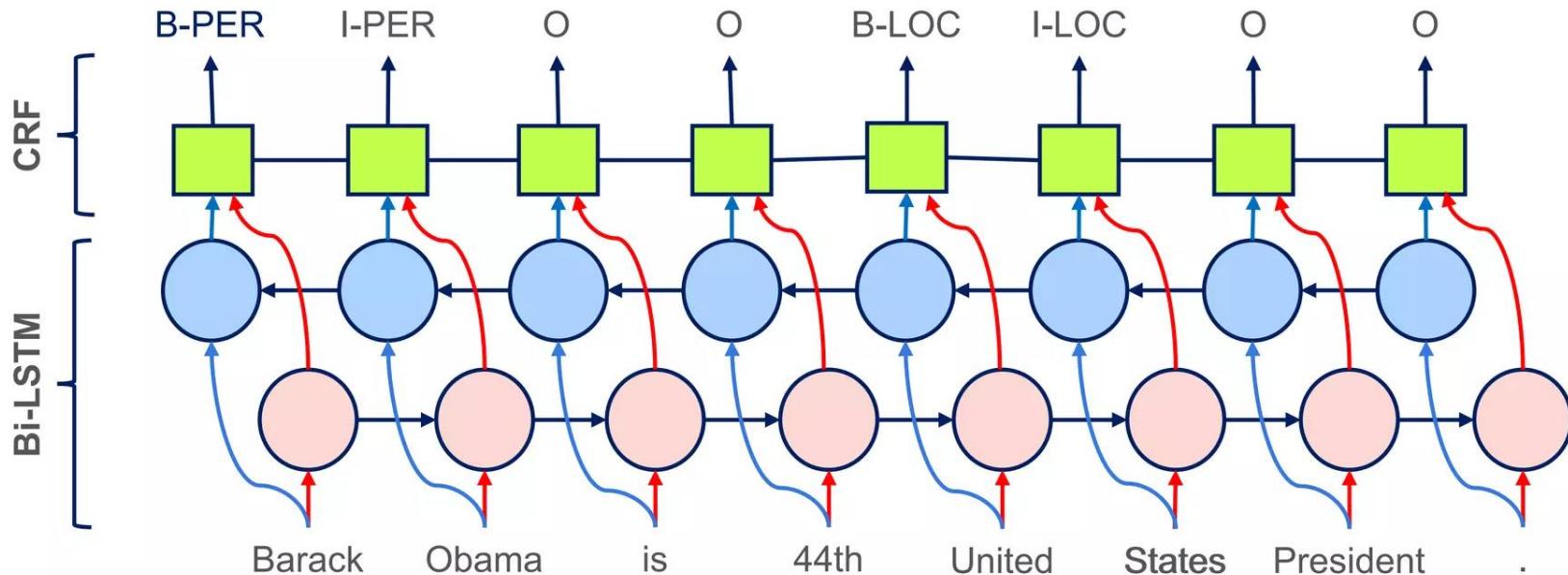
Neural Model - BiLSTM



- Input is sequence of tokens, output is sequence of BIO tags.
- Weights trained end-to-end, no feature engineering needed.
- Bidirectional LSTM gets signal from neighboring words on both sides.



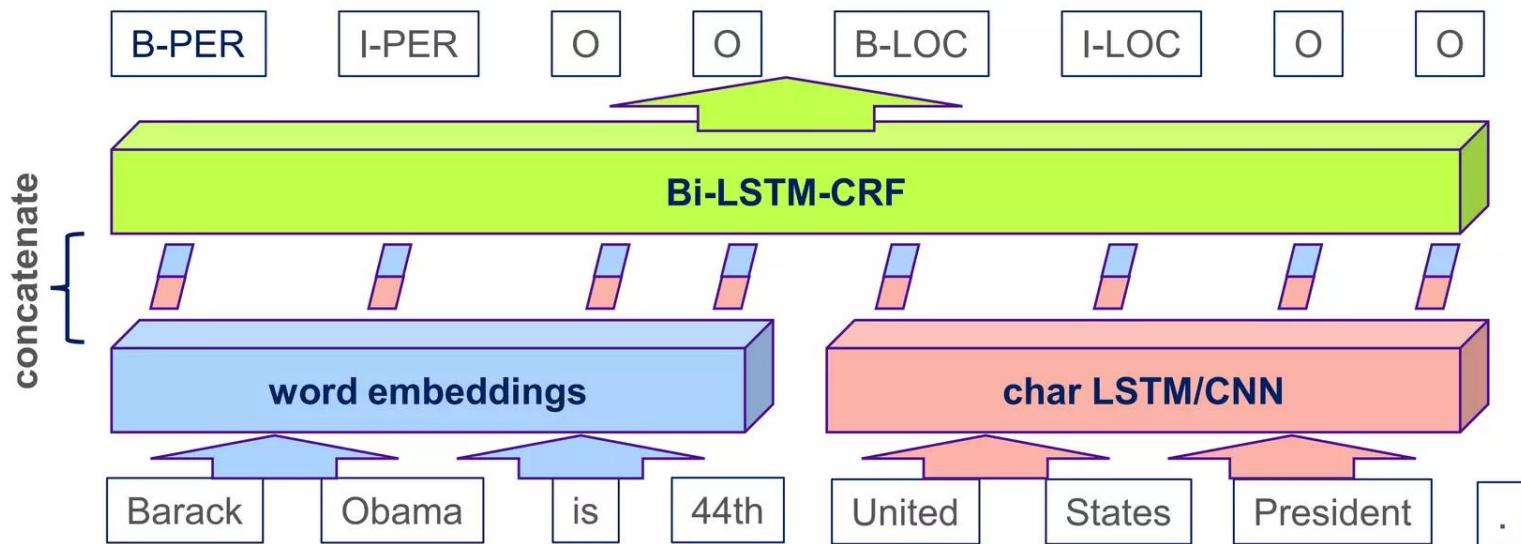
Neural Model – BiLSTM-CRF



- Same as previous model, with additional CRF layer.
- No feature engineering for CRF, unlike CRF only NER model.
- Pre-trained embeddings observed to improve performance.



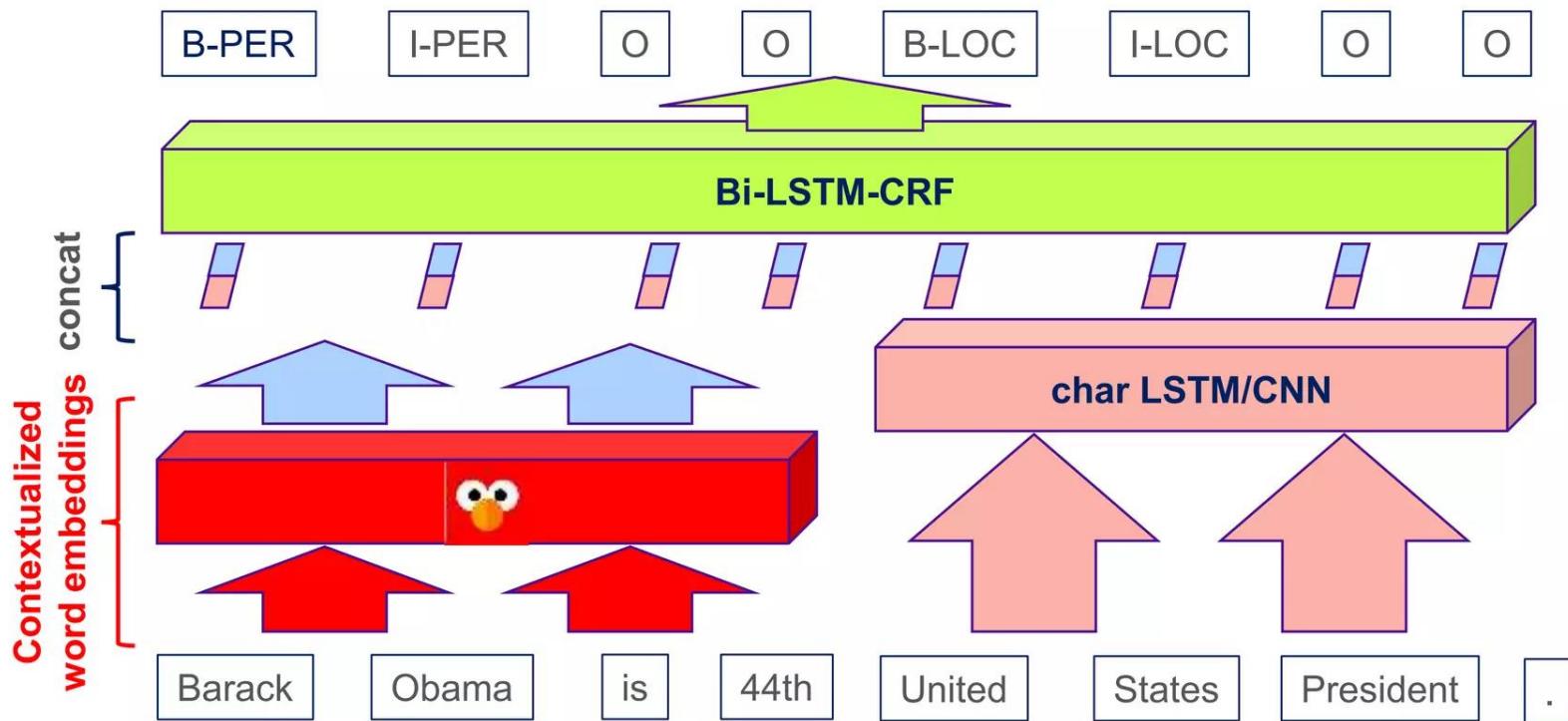
Neural Model – adding char embeddings



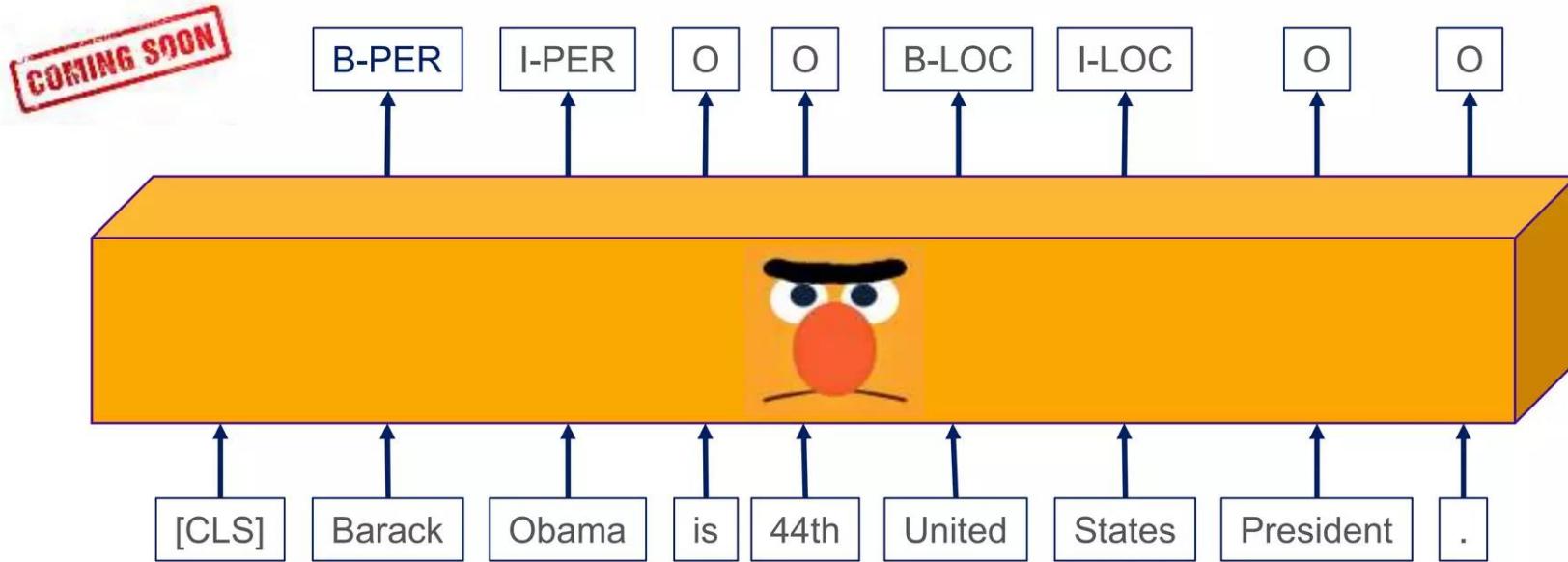
- Concatenate char embedding + word embedding and feed to Bi-LSTM-CRF.
- All weights learned end-to-end.
- Handles rare / unknown words; Exploits signal in prefix/suffix.



Neural Model – ELMo preprocessing



Neural Model – Transformer based



- BERT = Bidirectional Encoder Representation for Transformers.
- Source of embeddings similar to ELMo in standard BiLSTM + CRF models, OR
- Fine-tune LM backed NERs such as HuggingFace's BertForTokenClassification.



NERDS Overview

- Framework that provides easy to use NER capabilities to Data Scientists.
- Wraps various popular third party NER models.
- Extendable, new third party NER tools can be added as needed.
- Software Engineering tooling to boost Data Science productivity.
- Looking for support, bug reports, contributions, and ideas.



Benefits of Unification

- Consistent API – all models are subclasses of NERModel.

```
class NERModel(BaseEstimator, ClassifierMixin):  
    def fit(self, X, y): pass  
    def predict(self, X): pass  
    def save(self, dir_path): pass  
    def load(self, dir_path): pass
```

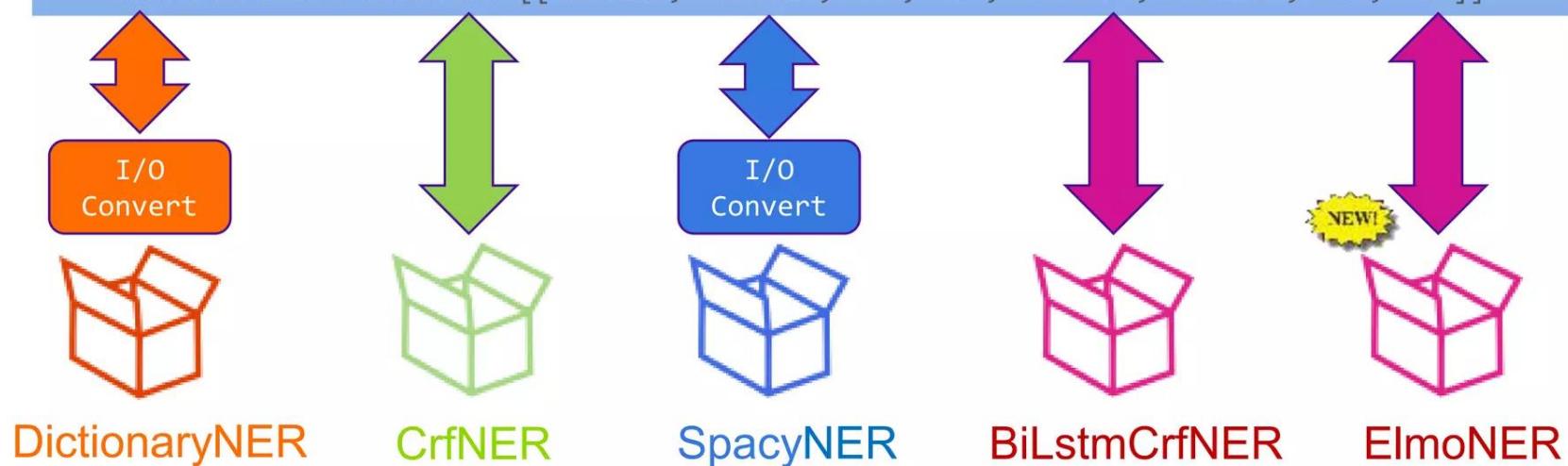
- Data prep. done once per project and reused across multiple models.
- Reusable Training and Evaluation code.
- Familiar Scikit-Learn like API, and access to Scikit-Learn utility functions.
- Duck-typing allows us to build Ensembles of NER.
- Easy to benchmark NER label data.



ELMo NER Model from Anago

Barack Obama is 44th United States President.
 PER LOC

Data: [[“Barack”, “Obama”, “is”, “44th”, “United” “States”, “President”, “.”]]
 Labels and Predictions: [[“B-PER”, “I-PER”, “0”, “0”, “B-LOC”, “I-LOC”, “0”, “0”]]



Ensemble NER

```
# create and test an ensemble
dict_model = DictionaryNER()
dict_model.load("models/dict_model")
crf_model = CrfNER()
crf_model.load("models/crf_model")
spacy_model = SpacyNER()
spacy_model.load("models/spacy_model")
bilstm_model = BiLstmCrfNER()
bilstm_model.load("models/bilstm_model")
model = EnsembleNER()
model.fit(xtrain, ytrain,
    estimators=[
        (dict_model, {}),
        (crf_model, {}),
        (spacy_model, {}),
        (bilstm_model, {})
    ],
    is_pretrained=True)
ypred = model.predict(xtest)
print(classification_report(flatten_list(ytest, strip_prefix=True),
                            flatten_list(ypred, strip_prefix=True),
                            labels=entity_labels))
```

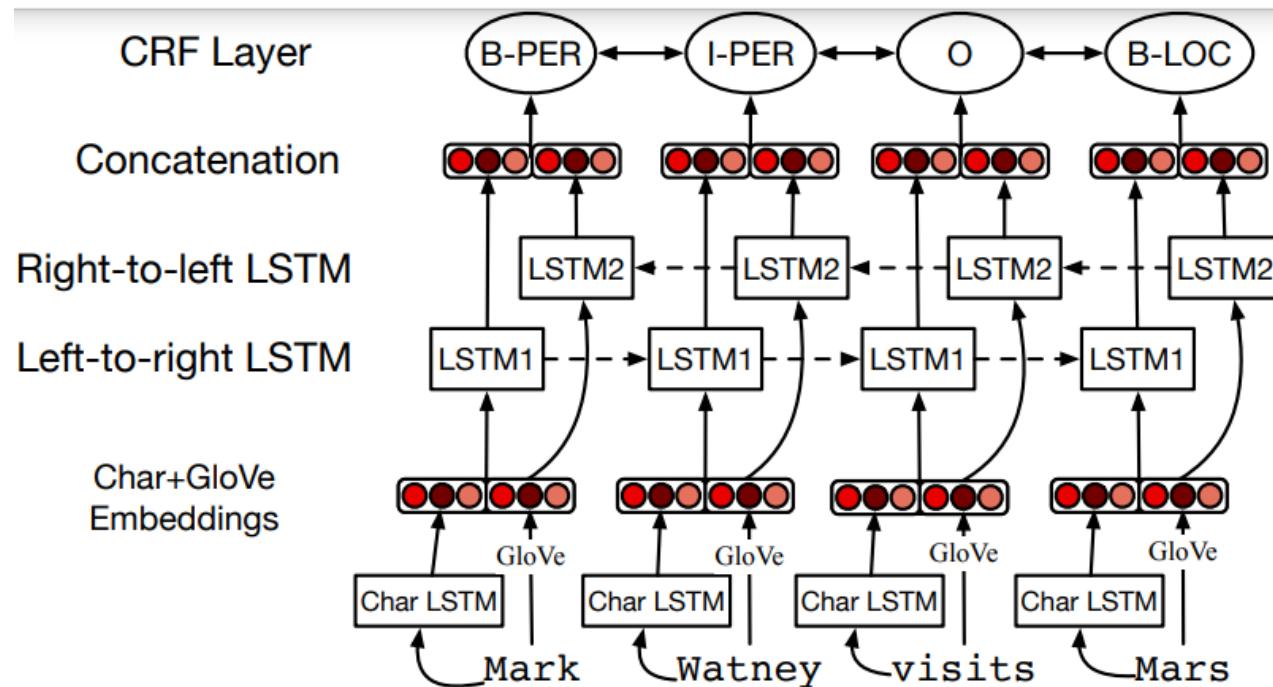
- Max Voting
- Improvements in this fork:
 - Unifies Max Voting and Weighted Max Voting NERs into single model.



Neural NER

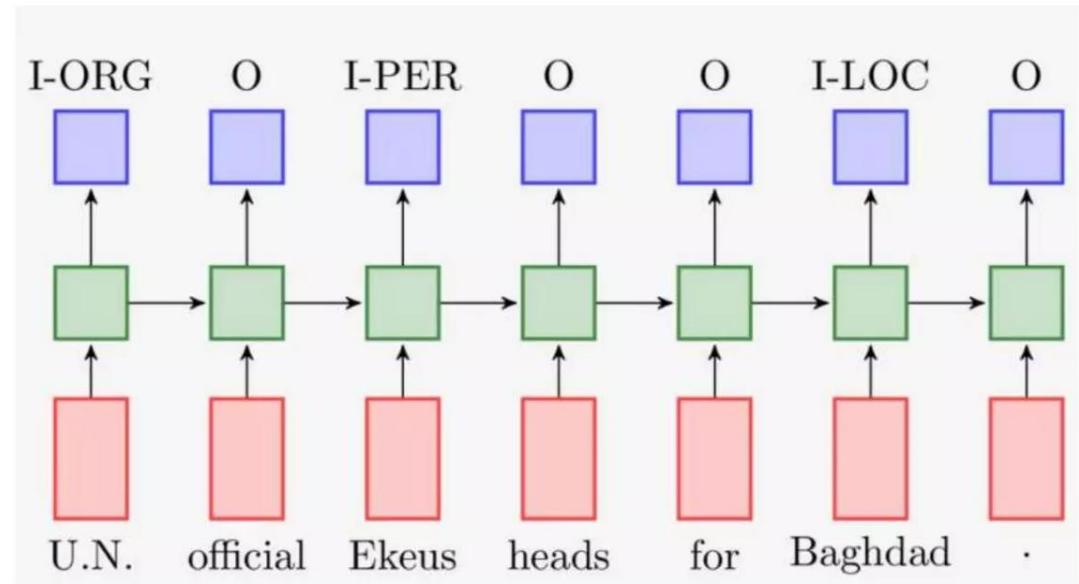
Sequence RNN (e.g. biLSTM or Transformer)
with a CRF output layer.

Input: word embeddings, possibly concatenated with character embeddings and other features, e.g.:



Neural NER

- Feature extraction?
- Embeddings
- LSTM



Evaluating IE Accuracy

- Always evaluate performance on independent, manually-annotated test data not used during system development.
- Measure for each test document:
 - Total number of correct extractions in the solution template: N
 - Total number of slot/value pairs extracted by the system: E
 - Number of extracted slot/value pairs that are correct (i.e. in the solution template): C
- Compute average value of metrics adapted from IR:
 - Recall = C/N
 - Precision = C/E
 - F-Measure = Harmonic mean of recall and precision

Slide by Chris Manning, based on slides by others

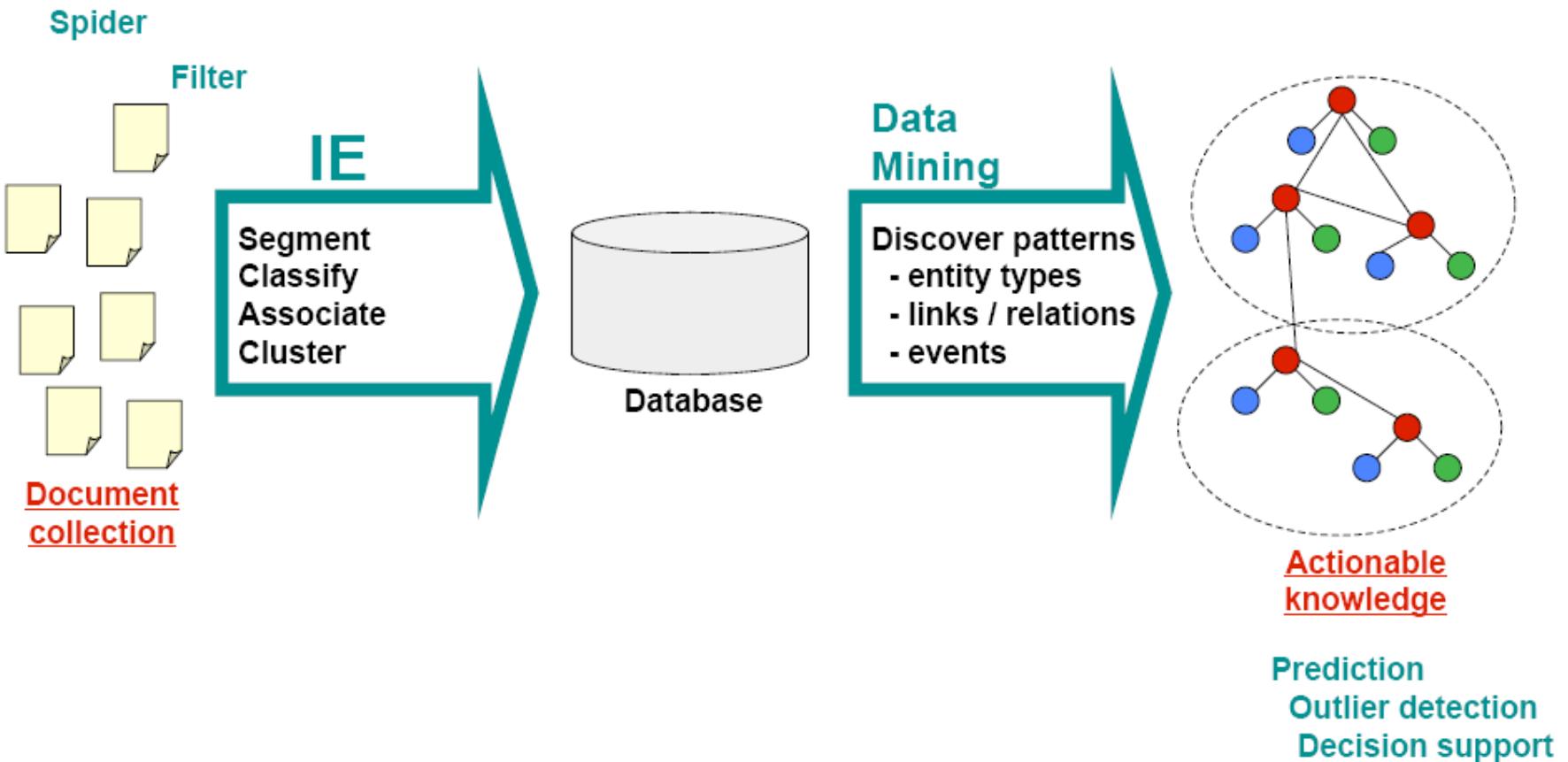
Precision, Recall, F1 for NER

- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funny for IE/NER when there are *boundary errors* (which are *common*):
 - First Bank of Chicago announced earnings ...
- This counts as both a fp and a fn
- Selecting *nothing* would have been better
- Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules)

MUC

- DARPA funded significant efforts in IE in the early to mid 1990's.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- Focused on extracting information from news articles:
 - Terrorist events
 - Industrial joint ventures
 - Company management changes
- Information extraction of particular interest to the intelligence community (CIA, NSA).

From Text to Actionable Knowledge



References

- <https://slideplayer.com/slide/3238232/>
- <https://www.youtube.com/watch?v=2-IK8TL2svo&t=8s>
- <https://www.youtube.com/watch?v=-I2gtDfqRJU>
- <https://www.youtube.com/watch?v=wxyZTSc2tM0>
- <https://www.youtube.com/watch?v=sm5ta8boAWY>

References

<https://demos.explosion.ai/displacy-ent>

[https://goodboychan.github.io/python/datacamp/natural_language_processing/2020/07/16/01-Named-entity-recognition.html\(jurafsky\)](https://goodboychan.github.io/python/datacamp/natural_language_processing/2020/07/16/01-Named-entity-recognition.html(jurafsky))

<https://slideplayer.com/slide/4235847/>

Named entity tagging

<https://www.youtube.com/watch?v=7CRyqwCZFY0>

NER

<https://www.bilibili.com/video/BV1CE41197rQ/?p=36>



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



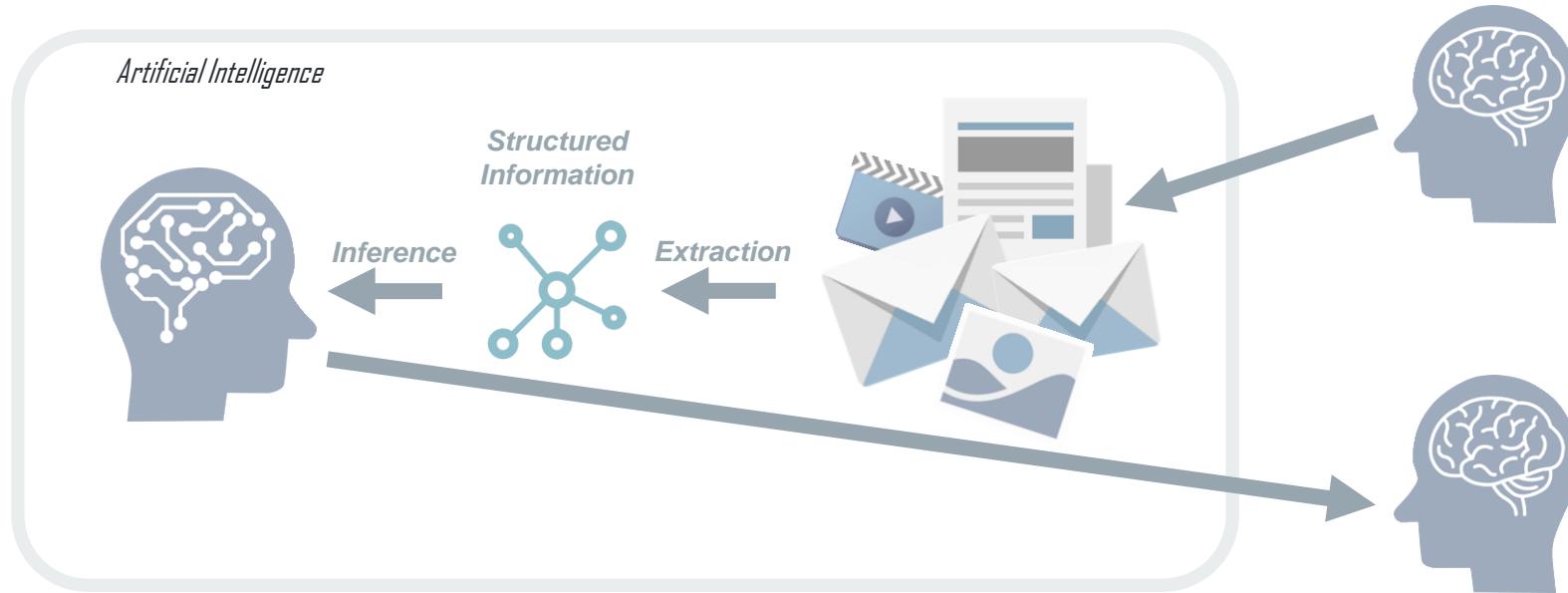


Session 13: Information Extraction-Relation Extraction

Date – 10th March 2024

These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philipp Koehn**, Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

A Quick Overview of Information Extraction

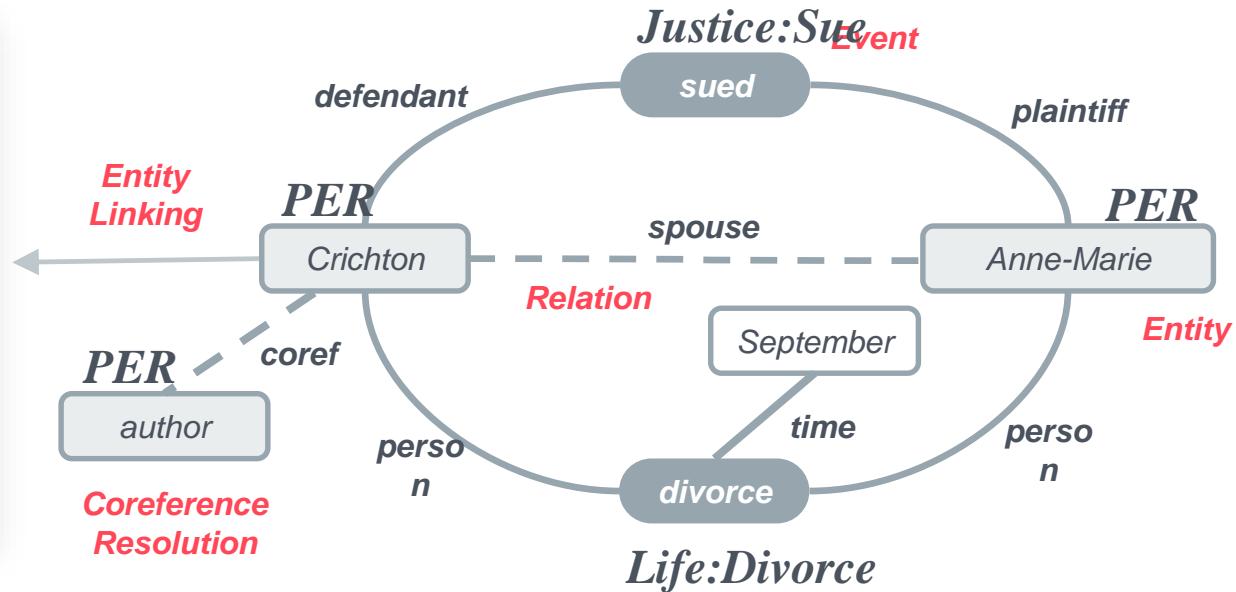


- With the exponential growth of data from various sources especially the Internet, there is an increasing need for **Information Extraction** technology that extracts **machine-readable structured information** to support downstream applications.

A Quick Overview of Information Extraction

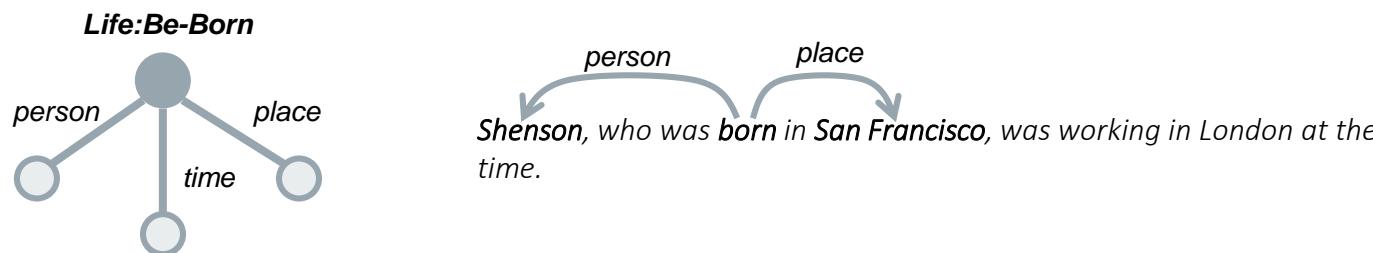
“

Anne-Marie sued Crichton, best known as the author of Jurassic Park, for divorce in September.



A Brief Introduction to Information Extraction Subtasks

- **Entity Extraction** aims to identify entity mentions in text and classify them into pre-defined entity types.
- **Relation Extraction** is the task of assigning a relation type to an ordered pair of entity mentions.
- **Event Extraction** entails identifying and classifying event triggers and their arguments
 - Event triggers: the words or phrase that most clearly express event occurrences
 - Arguments: the words or phrases for participants in those events



- **Entity Coreference Resolution** is the task of resolving all entity mentions that refer to the same entity.

Relation Extraction

- Up until now we have focused on early stages of the Information Extraction pipeline
 - We have emphasized named entity tagging
- Now we will discuss extracting facts about these entities
 - This can include IS-A facts (similar to named entity types), but also more complicated relations

Extracting relations from text

- **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)...”
- 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)

Extracting Relation Triples from Text



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox
Print/export

Languages
العربية
Azerbaijani
Беларуская
Беларуская (тарашкевіца)

Article Talk Read Edit View history Search

Stanford University

From Wikipedia, the free encyclopedia

Coordinates: 37.43°N 122.17°W

"Stanford" redirects here. For other uses, see [Stanford \(disambiguation\)](#).

Not to be confused with [Stamford University \(disambiguation\)](#).

The Leland Stanford Junior University, commonly referred to as **Stanford University** or **Stanford**, is an American private research university located in **Stanford**, California on an 8,180-acre (3,310 ha) campus near **Palo Alto**, California, United States. It is situated in the northwestern Santa Clara Valley on the San Francisco Peninsula, approximately 20 miles (32 km) northwest of **San Jose** and 37 miles (60 km) southeast of **San Francisco**.^[6]

Leland Stanford, a Californian railroad tycoon and politician, founded the university in 1891 in honor of his son, **Leland Stanford, Jr.**, who died of **typhoid** two months before his 16th birthday. The university was established as a coeducational and nondenominational institution, but struggled financially after the senior Stanford's 1893 death and after much of the campus was damaged by the 1906 San Francisco earthquake. Following World War II, Provost **Frederick Terman** supported faculty and graduates' entrepreneurialism to build a self-sufficient local industry in what would become known as **Silicon Valley**. By 1970, Stanford was home to a **linear accelerator**, was one of the original four **ARPANET** nodes, and had transformed itself into a major research university in computer science, mathematics, natural sciences, and social sciences. More than 50 Stanford faculty, staff, and alumni have won the **Nobel Prize** and Stanford has the largest number of Turing award winners for a single institution. Stanford faculty and alumni have founded many prominent technology companies including **Cisco Systems**, **Google**, **Hewlett-Packard**, **LinkedIn**, **Rambus**, **Silicon Graphics**, **Sun Microsystems**, **Varian Associates**, and **Yahoo!**^[7]

The university is organized into seven schools including academic schools of **Humanities**

The seal of Stanford University is circular with a red border. The outer ring contains the text "LELAND STANFORD JUNIOR UNIVERSITY" at the top and "DIE LUFT DER FREIHEIT WEHT" at the bottom. In the center is a redwood tree, with the year "1891" at the base.

rd Junior
only referred to as
or Stanford, is an
research university
California ... near
ia... Leland
the university in



Why Relation Extraction?

- Create new structured knowledge bases, useful for any app
- Augment current knowledge bases
- Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
- Support ques on answering

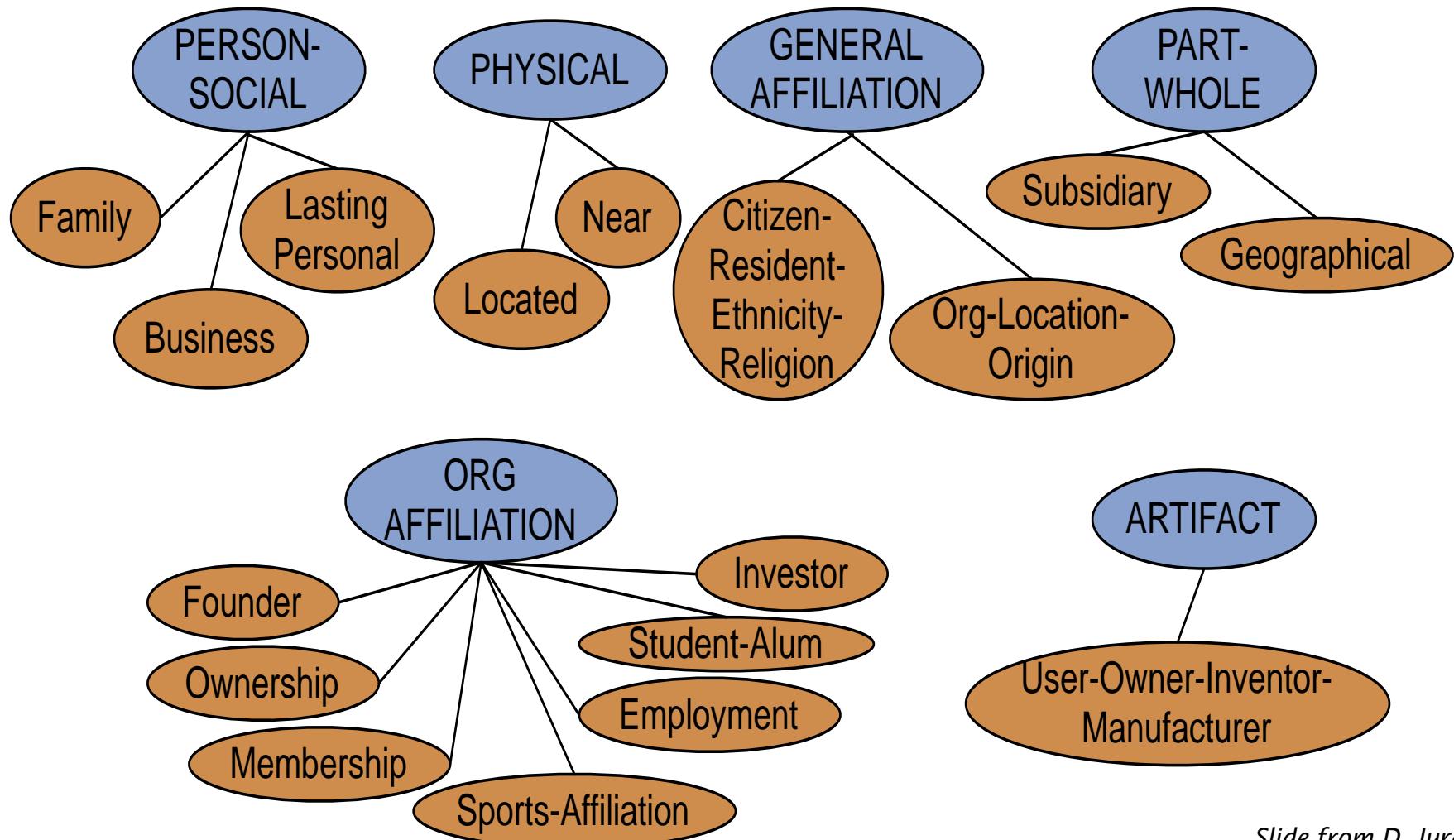
The granddaughter of which actor starred in the movie “E.T.”? (acted-in ?x “E.T.”)(is-a ?y actor)(granddaughter-of ?x ?y)

- But which relations should we extract?

Automated Content Extraction (ACE)



17 relations from 2008 “Relation Extraction Task”





Automated Content Extraction (ACE)

- 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...

UMLS: Unified Medical Language System



- 134 entity types, 54 relations

Injury

Bodily Location

Anatomical Structure

Pharmacologic Substance

Pharmacologic Substance

disrupts

location-of

part-of

causes

treats

Physiological Function

Biologic Function

Organism

Pathological Function

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

Databases of Wikipedia Relations



Wikipedia Infobox

```
 {{Infobox university
|image_name= Stanford University seal.svg
|image_size= 210px
|caption = Seal of Stanford University
|name =Stanford University
|native_name =Leland Stanford Junior University
|Motto = {{lang|de|"Die Luft der Freiheit weht"}}, r /> ([[German language|German]])<ref
|name="casper">{{cite speech|title=Die Luft der Freiheit weht—On and Off|author=Gerhard
Casper|first=Gerhard|last=Casper|authorlink=Gerhard Casper|date=1995-10-
|url=http://www.stanford.edu/dept/pres-provost/president/speeches/951005dieluft.html}}</ref>
|Mottoeng = The wind of freedom blows<ref name="casper" />
|established = 1891<ref>{{cite web |
|url=http://www.stanford.edu/home/stanford/history/begin.html | title=Stanford University History |
|publisher = Stanford University | accessdate = 2017-04-26}}</ref>
|type = [[private university|Private]]
|calendar= Quarter
|president = [[John L. Hennessy]]
|provost = [[John Etchemendy]]
|city = [[Stanford, California|Stanford]]
|state = California
|country = U.S.
```

Relations extracted from Infobox

Stanford state California

Stanford motto “Die Luft der Freiheit weht”

...

Relation databases that draw from Wikipedia

- 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,
people/person/profession,

biology/organism_higher_classification

location/location/contains
people/person/place-of-birth

film/film/genre

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

Patterns for Relation Extraction



- Hand-written rules for relation extraction were used in MUC (such as the Fastus system)
- Recently there has been a renewed wide interest in learning rules for relation extraction focused on precision
 - The presumption is that interesting information occurs many times on the web, with different contexts
 - e.g., how many times does "Barack Obama is the 44th President of the United States" occur on the web?
 - Focusing on high precision is reasonable because the high redundancy will allow us to deal with recall

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
 4. Deep Learning
-

Rules for extracting IS-A relation



Early intuition from Hearst (1992)

- “Agar is a substance prepared from a mixture of red algae, such as *Gelidium*, for laboratory or industrial use”
- What does *Gelidium* mean?
- How do you know?`

Rules for extracting IS-A relation



Early intuition from Hearst (1992)

- “Agar is a substance prepared from a mixture of red algae, such as **Gelidium**, for laboratory or industrial use”
- What does *Gelidium* mean?
- How do you know?`

Hearst's Patterns for extracting IS-A relations



(Hearst, 1992): *Automatic Acquisition of Hyponyms*

"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"

Hearst's Patterns for extracting IS-A relations

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...

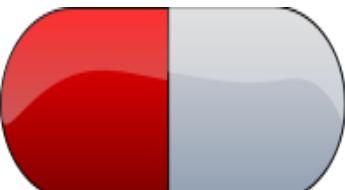
Extracting Richer Relations Using Rules



- 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!

Which relations hold between 2 entities?

Named Entities aren't quite enough.



Drug

Cure?

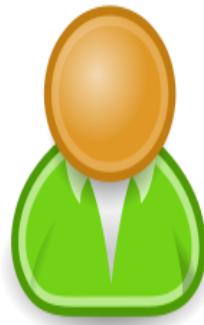
Prevent?

Cause?



Disease

What relations hold between 2 entities?



PERSON

Founder?

Investor?

Member?

Employee?

President?



ORGANIZATION

Extracting Richer Relations Using Rules and Named Entities

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

Idea: define some extraction patterns

X is the founder of Y
X, who founded Y
Y was founded by X



48-year-old Elon Musk is the founder of SpaceX and a co-founder of Tesla Motors.
Elon Musk, who founded SpaceX in 2002, has said the company is focused on ...
SpaceX was founded by Elon Musk to make life multi-planetary. "You want to ..."

Problem: most occurrences do not fit simple patterns

You may also be thinking of Elon Musk (founder of SpaceX), who started PayPal.
Elon Musk, co-founder of PayPal, went on to establish SpaceX, one of the most ...
If Space Exploration (SpaceX), founded by Paypal pioneer Elon Musk succeeds, ...

Stanford



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

Supervised Methods

- For named entity tagging, statistical taggers are the state of the art
- However, for relation extraction, this is not necessarily true
 - Still many hand-crafted rule-based systems out there that work well
 - But hand-crafting such systems takes a lot of work, so classification approaches are very interesting (and they are improving with time)
- Formulate relation extraction as a supervised classification problem

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

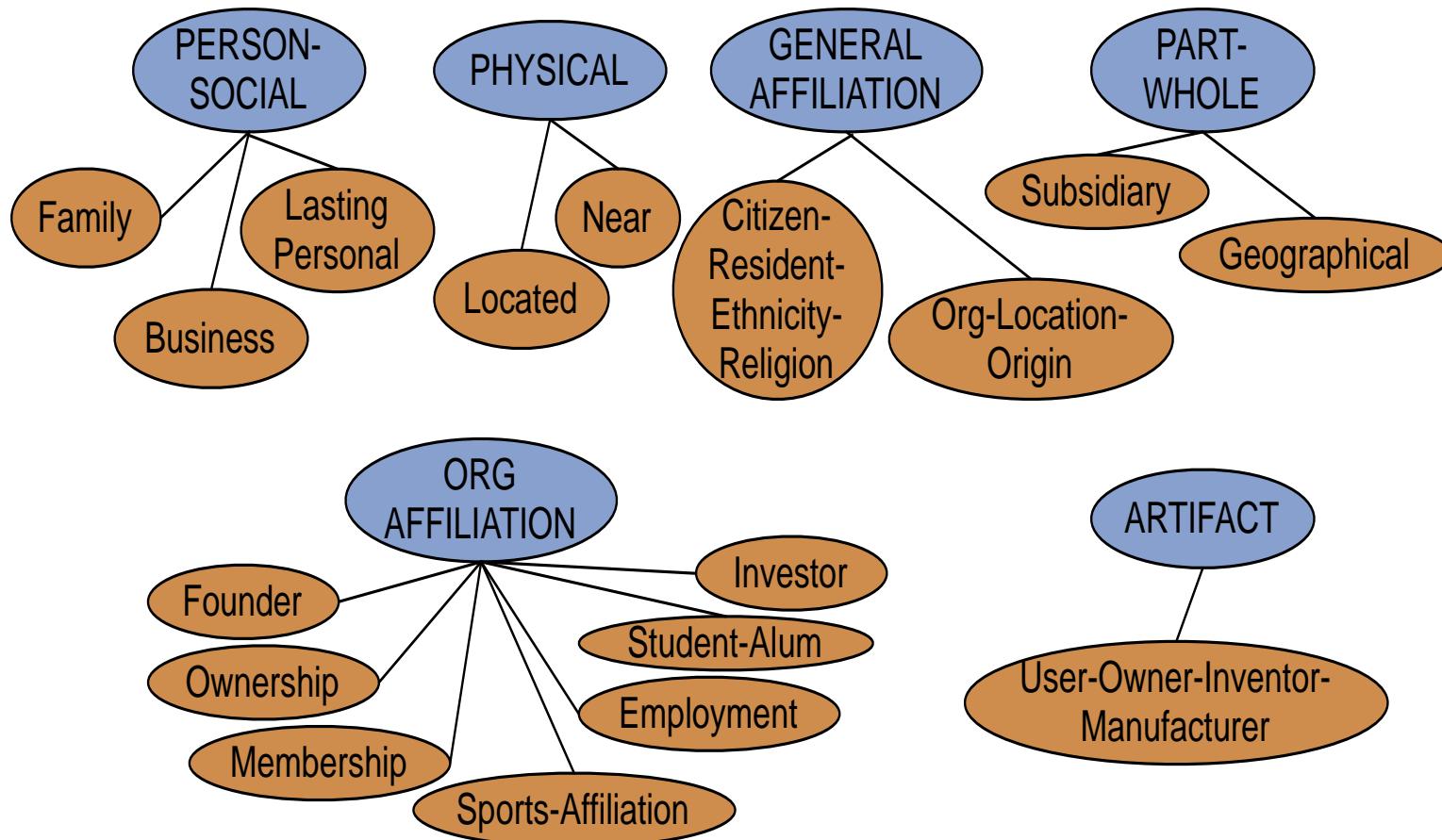
How to do classification in supervised relation extraction

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.

Automated Content Extraction (ACE)



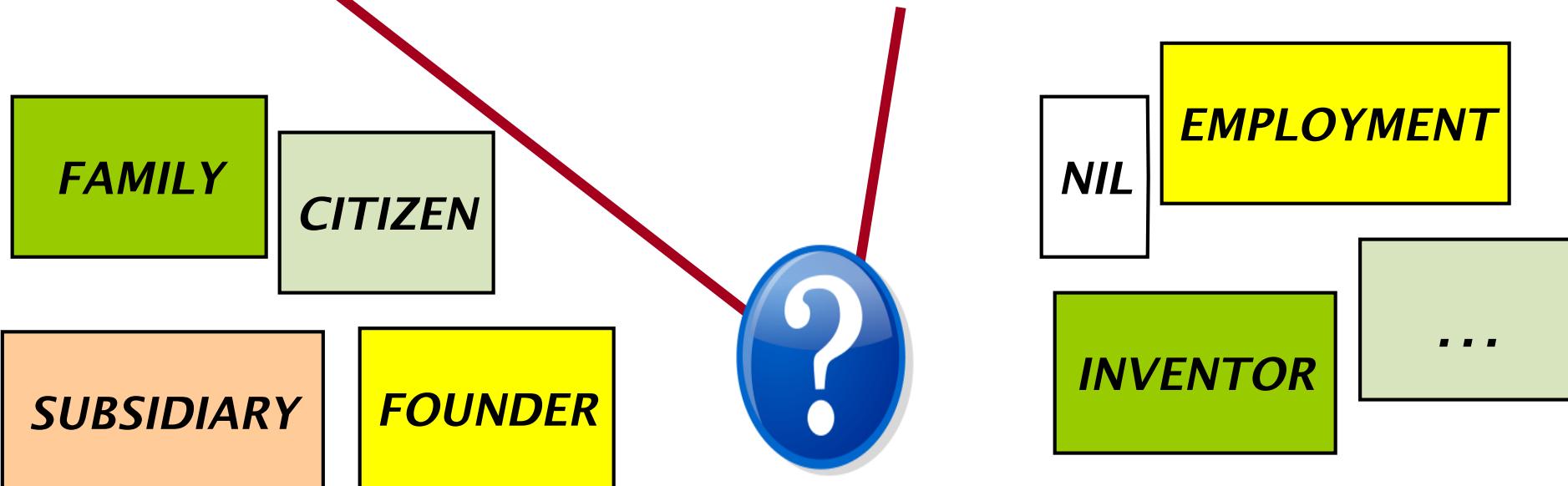
*17 sub-relations of 6 relations from 2008
“Relation Extraction Task”*



Relation Extraction

Classify the relation between two entities in a sentence

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



Slide from D. Jurafsky

Word Features for Relation Extraction



Mention 1

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

Mention 2

- 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}

Named Entity Type and Mention Level Features for Relation Extraction

Mention 1

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

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]

Parse Features for Relation Extraction

Mention 1

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

Mention 2

- 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

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

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>

Classifiers for supervised methods

- Now you can use any classifier you like
 - Decision Tree
 - MaxEnt
 - Naïve Bayes
 - SVM
 - ...
- Train it on the training set, tune on the dev set, test on the test set

Idea: label examples, train a classifier



Success! Better generalizability

Problem: labeling examples is expensive :-(

Stanford



Summary: Supervised Relation Extraction

- + Can get high accuracies with enough hand-labeled training data, if test similar enough to training
- Labeling a large training set is expensive
- Supervised models are brittle, don't generalize well to different genres

Semi-Supervised Methods

- We'd like to minimize our reliance on having a large training set
- Instead, given a few examples or a few high-precision patterns, we'd like to generalize
 - This is sometimes referred to as "bootstrapping"

Relation Bootstrapping (Hearst 1992)

- Gather a set of seed pairs that have relation R
- Iterate:
 1. Find sentences with these pairs
 2. Look at the context between or around the pair and generalize the context to create patterns
 3. Use the patterns to grep for more pairs

Bootstrapping

- <Mark Twain, Elmira> **Seed tuple**
 - Grep (google) for the environments of the seed tuple
 - “Mark Twain is buried in Elmira, NY.”
 - X is buried in Y
 - “The grave of Mark Twain is in Elmira”
 - The grave of X is in Y
 - “Elmira is Mark Twain’s final resting place”
 - Y is X’s final resting place.
- Use those patterns to grep for new tuples
- Iterate

Dipre:

Extract <author, book> pairs

Brin, Sergei. 1998. Extracting Patterns and Relations from the World Wide Web.

- Start with 5 seeds:

Author	Book
Isaac Asimov	The Robots of Dawn
David Brin	Startide Rising
James Gleick	Chaos: Making a New Science
Charles Dickens	Great Expectations
William Shakespeare	The Comedy of Errors

- Find Instances:

The Comedy of Errors, by William Shakespeare, was

The Comedy of Errors, by William Shakespeare, is

The Comedy of Errors, one of William Shakespeare's earliest attempts

The Comedy of Errors, one of William Shakespeare's most

- Extract patterns (group by middle, take longest common prefix/suffix)

?x , by ?y ,

?x , one of ?y 's

- Now iterate, finding new seeds that match the pattern

Snowball

- Similar iterative algorithm

Organization	Location of Headquarters
Microsoft	Redmond
Exxon	Irving
IBM	Armonk

- Group instances w/similar prefix, middle, suffix, extract patterns
 - But require that X and Y be named entities
 - And compute a confidence for each pattern

.69 *ORGANIZATION* { 's, in, headquarters} *LOCATION*

.75 *LOCATION* {in, based} *ORGANIZATION*

Distant Supervision

- Combine bootstrapping with supervised learning
 - Instead of 5 seeds,
 - Use a large database to get huge # of seed examples
 - Create lots of features from all these examples
 - Combine in a supervised classifier

Snow, Jurafsky, Ng. 2005. *Learning syntactic patterns for automatic hypernym discovery*. NIPS 17

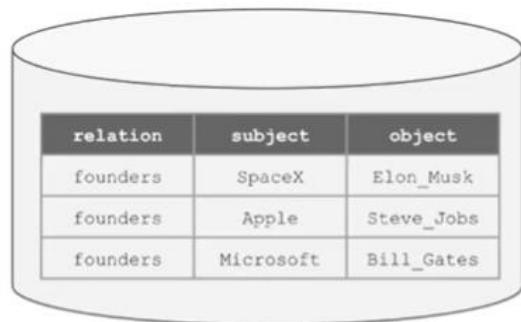
Fei Wu and Daniel S. Weld. 2007. *Autonomously Semantifying Wikipedia*. CIKM 2007

Mintz, Bills, Snow, Jurafsky. 2009. *Distant supervision for relation extraction without labeled data*. ACL09

Idea: derive labels from an existing knowledge base (KB)

Assume sentences with related entities are positive examples

Assume sentences with unrelated entities are negative examples



Elon Musk, co-founder of PayPal, went on to establish SpaceX, one of the most ...	<input checked="" type="checkbox"/>
Entrepreneur Elon Musk announced the latest addition to the SpaceX arsenal ...	<input checked="" type="checkbox"/>
Elon Musk dismissed concerns that Apple was poaching the company's talent.	<input checked="" type="checkbox"/>
Now we know what Apple would have done with Elon Musk if that deal had ...	<input checked="" type="checkbox"/>

Hooray! Massive quantities of training data, practically free!

Qualm: are those assumptions reliable?

Stanford





Distant supervision paradigm

- Like supervised classification:
 - Uses a classifier with lots of features
 - Supervised by detailed hand-created knowledge
 - Doesn't require iteratively expanding patterns
- Like unsupervised classification:
 - Uses very large amounts of unlabeled data
 - Not sensitive to genre issues in training corpus

Distantly supervised learning of relation extraction patterns

- 1 For each relation
- 2 For each tuple in big database
- 3 Find sentences in large corpus with both entities
- 4 Extract frequent features (parse, words, etc)
- 5 Train supervised classifier using thousands of patterns

Born-In

<Edwin Hubble, Marshfield>

<Albert Einstein, Ulm>

Hubble was born in Marshfield

Einstein, born (1879), Ulm

Hubble's birthplace in Marshfield

PER was born in LOC

PER, born (XXXX), LOC

PER's birthplace in LOC

$P(\text{born-in} / f_1, f_2, f_3, \dots, f_{70000})$

Distant supervision- Supervised Approach



Distant supervision is a powerful idea — but it has two limitations:

1. Not all sentences with related entities are truly positive examples

Entrepreneur Elon Musk announced the latest addition to the SpaceX arsenal ...



(but the benefit of *more* data outweighs the harm of noisier data)

2. Need an existing KB to start from — can't start from scratch



Unsupervised relation extraction



- Open Information Extraction:
 - extract relations from the web with no training data, no list of relations
1. Use parsed data to train a “trustworthy tuple” classifier
 2. Single-pass extract all relations between NPs, keep if trustworthy
 3. Assessor ranks relations based on text redundancy
 - (FCI, specializes in, software development)
 - (Tesla, invented, coil transformer)

M. Banko, M. Cararella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. IJCAI

Evaluation of Supervised Relation Extraction

- Compute P/R/F₁ for each relation

$$P = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of extracted relations}}$$

$$F_1 = \frac{2PR}{P + R}$$

$$R = \frac{\text{\# of correctly extracted relations}}{\text{Total \# of gold relations}}$$

Evaluation of Semi-supervised and Unsupervised Relation Extraction

- Since it extracts totally new relations from the web
 - There is no gold set of correct instances of relations!
 - Can't compute precision (don't know which ones are correct)
 - Can't compute recall (don't know which ones were missed)
- Instead, we can approximate precision (only)
 - Draw a random sample of relations from output, check precision manually

$$\hat{P} = \frac{\text{\# of correctly extracted relations in the sample}}{\text{Total \# of extracted relations in the sample}}$$

- Can also compute precision at different levels of recall.
 - Precision for top 1000 new relations, top 10,000 new relations, top 100,000
 - In each case taking a random sample of that set
- But no way to evaluate recall

Method	Input	Output	Description	Advantages	Disadvantages	innovate	achieve	lead
Bootstrapping	Unlabeled text, relation schema, rules and/or examples	Extraction rules, relations	Using a small set of extraction rules, extract examples, keep prominent ones, iteratively learn more extraction rules and examples	Easy to add new rules, can also be supplied by user	Often low recall and/or manual refinement needed for high precision			
Rule-based	Unlabeled text, relation schema, rules and NE gazetteers	Relations	Using extraction rules and gazetteers, extract relations	Easy to add new rules, can also be supplied by user	Often low recall, much manual effort to develop			
Supervised	Labeled text, relation schema	Relations	Using a schema and labeled training data, train model	Currently highest precision and recall for schema-specific relation extraction	Up-front effort of labeling data, risk of overfitting training set			
Open IE	Unlabeled text	Groups of relations	Discover groups of relations from text using clustering, keep prominent ones	No knowledge about text needed	Difficult to make sense of groups and map to relation schemas			
Distantly Supervised	Unlabeled text, relation schema, examples	Extraction model, relations	Using a schema and examples of relations, automatically annotate training data, train a model to extract more relations	Extracting relations with high recall and precision	Initial examples required			
Universal Schema	Several partly populated knowledge bases	Unified knowledge	Take several KBs defined by different schemas, partly populated with relations, predict union of KBs	Integrate relations defined by different schemas after extraction	For small KBs it can be faster to do this manually			

Distant Supervision

- Mark joined **Amazon** a month ago.
 - What is the entity type?
- Weak Supervision:
 - From knowledge bases
 - Amazon.com, Inc is an American multinational technology company.

Supervision Sources for IE

- Mark joined **Amazon** a month ago.
- Direct Supervision: seen similar examples
- Weak Supervision:
 - From knowledge bases: **Amazon.com, Inc** is an American multinational technology company.
 - From linguistic patterns: PER *join* company
 - From pre-trained LMs: Amazon is a [MASK] <- [MASK] = company.
 - From task and label definitions: company is an organization that...
 - From global statistics and biases
- Indirect Supervision:
 - Entailment model: Amazon is a company. <- entailment
 - QA model: Q: What is Amazon? A: a company (that Mark joined).
 - Summarization model: Mark joined a new company.

Distant Supervision

- Weak Supervision (aka. Distant Supervision)
 - Directly related to the task
 - Do not cost additional annotation
 - Often noisy (i.e., are not completely accurate, or do not cover all aspects)
- Common Sources
 - Existing “free-to-use” knowledge bases, databases, dictionaries
 - Unannotated free-form texts (LM pre-training)
 - Linguistic patterns and templates (open IE)
 - General rules and minimum human inputs (label definitions)

Distant Supervision – Knowledge Bases

- One of the earliest attempts: entity and entity relations
- Hoffmann et al. (2010): Learn from Wikipedia infoboxes

Personal details	
Born	November 19, 1949 (age 72) New York City, New York, U.S.
Spouse(s)	Michael W. Doyle (m. 1976)
Children	Abigail
Education	Harvard University (BA, PhD) London School of Economics (MS)

Amy Gutmann was born on November 19, 1949, [2] in Brooklyn, New York, [2] the only child of Kurt and Beatrice Gutmann. ... She then entered Radcliffe College of Harvard University in 1967 on a scholarship as a math major with sophomore standing. ... She and her husband Michael Doyle have also funded an endowed undergraduate scholarship and an undergraduate research fund at Penn.

- Matching info box entities with context, to learn context-dependent relation extraction.
 - 5000+ relations
- Many follow-up work on de-noising, but with similar weak signals

Distant Supervision

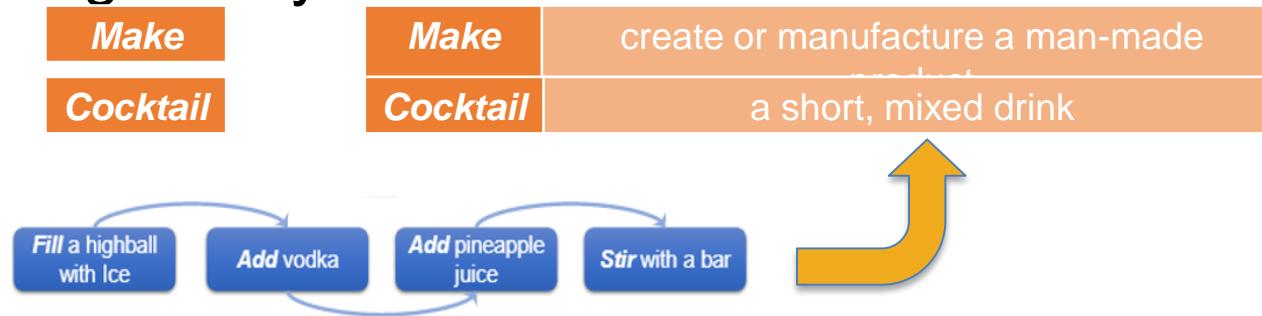


- Mark joined **Amazon** a month ago.
 - What is the entity type?
- Weak Supervision:
 - From knowledge bases 
 - Amazon.com, Inc is an American multinational technology company.
 - From weak but richer label representations
 - Word-embedding(company) is close to Word-embedding(Amazon)

Distant Supervision

– Label Representations

- Chen et al. (2020): Event Process Typing
- Direct label understanding is difficult
 - Add glossary definition as a “weak” label definition



Why using label glosses?

- Semantically richer than labels themselves
- Capturing the association of a process-gloss pair (two sequences) is much easier
- Jump-starting few-shot label representations (and benefiting with fairer prediction)

Chen et al. “What Are You Trying to Do?” Semantic Typing of Event Processes. CoNLL 2020

Distant Supervision Approaches

- Mark joined **Amazon** a month ago.
 - What is the entity type?
- Weak Supervision:
 - From knowledge bases
 - Amazon.com, Inc is an American multinational technology company.
 - From weak but richer label representations
 - Word-embedding(company) is close to Word-embedding(Amazon)
 - From pre-trained LMs
 - Amazon is a [MASK] <- [MASK] = company

Distant Supervision

– Pretrained Language Models

- Pre-trained language models can also be used as distant supervision
 - It did not use additional annotations
 - It is not task-specific
 - It contains inductive biases (weak signals)
- PLMs are applied for IE in many creative ways
 - Contextual embeddings to replace word embeddings
 - Direct probing
 - Direct probing + task-specific finetuning

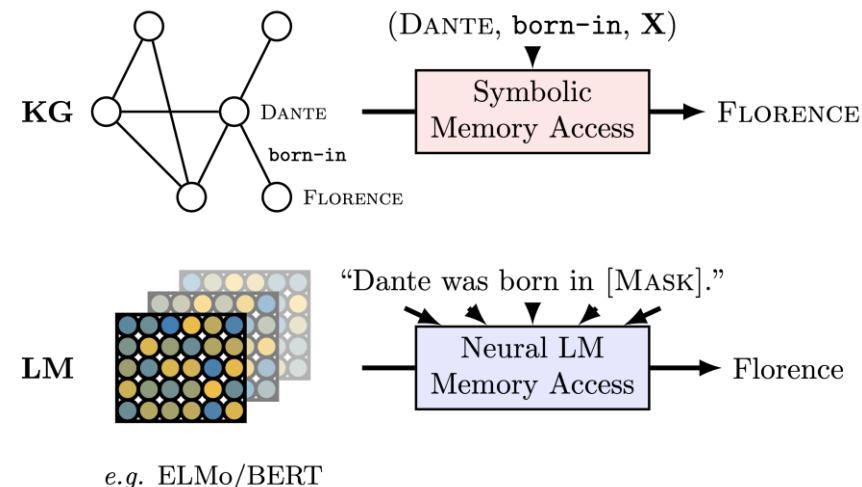
Paris is a
[MASK].



Distant Supervision

– Pretrained Language Models

- Comparing to ELMo, BERT made direct probing easier
- Petroni et al. (2019): Language models as knowledge bases
 - Google-RE
 - 16.1% birth-place
 - 1.4% birth-date



Petroni et al. *Language Models as Knowledge Bases?*. EMNLP 2019

Distant Supervision

– Pretrained Language Models

- Petroni et al. (2019): Language models as knowledge bases

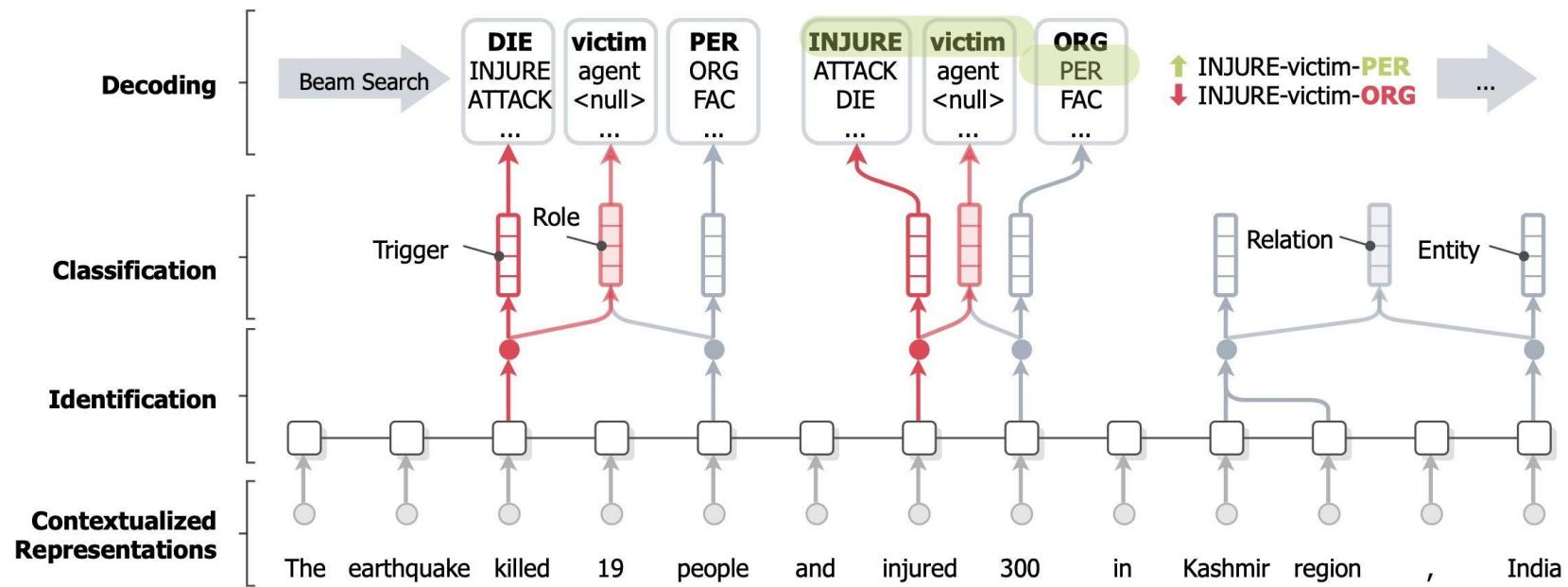
Relation	Query	Answer	Generation	
P19	Francesco Bartolomeo Conti was born in ____.	Florence	Rome [-1.8], Florence [-1.8], Naples [-1.9], Milan [-2.4], Bologna [-2.5]	
P20	Adolphe Adam died in ____.	Paris	Paris [-0.5], London [-3.5], Vienna [-3.6], Berlin [-3.8], Brussels [-4.0]	
P279	English bulldog is a subclass of ____.	dog	dogs [-0.3], breeds [-2.2], dog [-2.4], cattle [-4.3], sheep [-4.5]	
P37	The official language of Mauritius is ____.	English	English [-0.6], French [-0.9], Arabic [-6.2], Tamil [-6.7], Malayalam [-7.0]	
P413	Patrick Oboya plays in ____ position.	midfielder	centre [-2.0], center [-2.2], midfielder [-2.4], forward [-2.4], midfield [-2.7]	
P138	Hamburg Airport is named after ____.	Hamburg	Hess [-7.0], Hermann [-7.1], Schmidt [-7.1], Hamburg [-7.5], Ludwig [-7.5]	
P364	The original language of Mon Oncle Benjamin is ____.	French	French [-0.2], Breton [-3.3], English [-3.8], Dutch [-4.2], German [-4.9]	
P54	Dani Alves plays with ____.	Barcelona	Santos [-2.4], Porto [-2.5], Sporting [-3.1], Brazil [-3.3], Portugal [-3.7]	
P106	Paul Toungui is a ____ by profession.	politician	lawyer [-1.1], journalist [-2.4], teacher [-2.7], doctor [-3.0], physician [-3.7]	
P527	Sodium sulfide consists of ____.	sodium	water [-1.2], sulfur [-1.7], sodium [-2.5], zinc [-2.8], salt [-2.9]	
P102	Gordon Scholes is a member of the ____ political party.	Labor	Labour [-1.3], Conservative [-1.6], Green [-2.4], Liberal [-2.9], Labor [-2.9]	
P530	Kenya maintains diplomatic relations with ____.	Uganda	India [-3.0], Uganda [-3.2], Tanzania [-3.5], China [-3.6], Pakistan [-3.6]	
P176	iPod Touch is produced by ____.	Apple	Apple [-1.6], Nokia [-1.7], Sony [-2.0], Samsung [-2.6], Intel [-3.1]	
P30	Bailey Peninsula is located in ____.	Antarctica	Antarctica [-1.4], Bermuda [-2.2], Newfoundland [-2.5], Alaska [-2.7], Canada [-3.1]	
P178	JDK is developed by ____.	Oracle	IBM [-2.0], Intel [-2.3], Microsoft [-2.5], HP [-3.4], Nokia [-3.5]	
P1412	Carl III used to communicate in ____.	Swedish	German [-1.6], Latin [-1.9], French [-2.4], English [-3.0], Spanish [-3.0]	
P17	Sunshine Coast, British Columbia is located in ____.	Canada	Canada [-1.2], Alberta [-2.8], Yukon [-2.9], Labrador [-3.4], Victoria [-3.4]	
P39	Pope Clement VII has the position of ____.	pope	cardinal [-2.4], Pope [-2.5], pope [-2.6], President [-3.1], Chancellor [-3.2]	
P264	Joe Cocker is represented by music label ____.	Capitol	EMI [-2.6], BMG [-2.6], Universal [-2.8], Capitol [-3.2], Columbia [-3.3]	
P276	London Jazz Festival is located in ____.	London	London [-0.3], Greenwich [-3.2], Chelsea [-4.0], Camden [-4.6], Stratford [-4.8]	
P127	Border TV is owned by ____.	ITV	Sky [-3.1], ITV [-3.3], Global [-3.4], Frontier [-4.1], Disney [-4.3]	
P103	The native language of Mammootty is ____.	Malayalam	Malayalam [-0.2], Tamil [-2.1], Telugu [-4.8], English [-5.2], Hindi [-5.6]	
P495	The Sharon Cuneta Show was created in ____.	Philippines	Manila [-3.2], Philippines [-3.6], February [-3.7], December [-3.8], Argentina [-4.0]	
ConceptNet	AtLocation	You are likely to find a overflow in a ____.	drain	sewer [-3.1], canal [-3.2], toilet [-3.3], stream [-3.6], drain [-3.6]
	CapableOf	Ravens can ____.	fly	fly [-1.5], fight [-1.8], kill [-2.2], die [-3.2], hunt [-3.4]
	CausesDesire	Joke would make you want to ____.	laugh	cry [-1.7], die [-1.7], laugh [-2.0], vomit [-2.6], scream [-2.6]
	Causes	Sometimes virus causes ____.	infection	disease [-1.2], cancer [-2.0], infection [-2.6], plague [-3.3], fever [-3.4]
	HasA	Birds have ____.	feathers	wings [-1.8], nests [-3.1], feathers [-3.2], died [-3.7], eggs [-3.9]
	HasPrerequisite	Typing requires ____.	speed	patience [-3.5], precision [-3.6], registration [-3.8], accuracy [-4.0], speed [-4.1]
	HasProperty	Time is ____.	finite	short [-1.7], passing [-1.8], precious [-2.9], irrelevant [-3.2], gone [-4.0]
	MotivatedByGoal	You would celebrate because you are ____.	alive	happy [-2.4], human [-3.3], alive [-3.3], young [-3.6], free [-3.9]
	ReceivesAction	Skills can be ____.	taught	acquired [-2.5], useful [-2.5], learned [-2.8], combined [-3.9], varied [-3.9]
	UsedFor	A pond is for ____.	fish	swimming [-1.3], fishing [-1.4], bathing [-2.0], fish [-2.8], recreation [-3.1]

These predictions are highly relevant to typing and relation extraction

- Mark joined **Amazon** a month ago.
 - What is the entity type?
- Weak Supervision:
 - From knowledge bases
 - Amazon.com, Inc is an American multinational technology company.
 - From weak but richer label representations
 - Word-embedding  company is close to Word-embedding(Amazon)
 - From pre-trained LMs
 - Amazon is a [MASK] <- [MASK] = company.
 - From linguistic patterns
 - PER join company

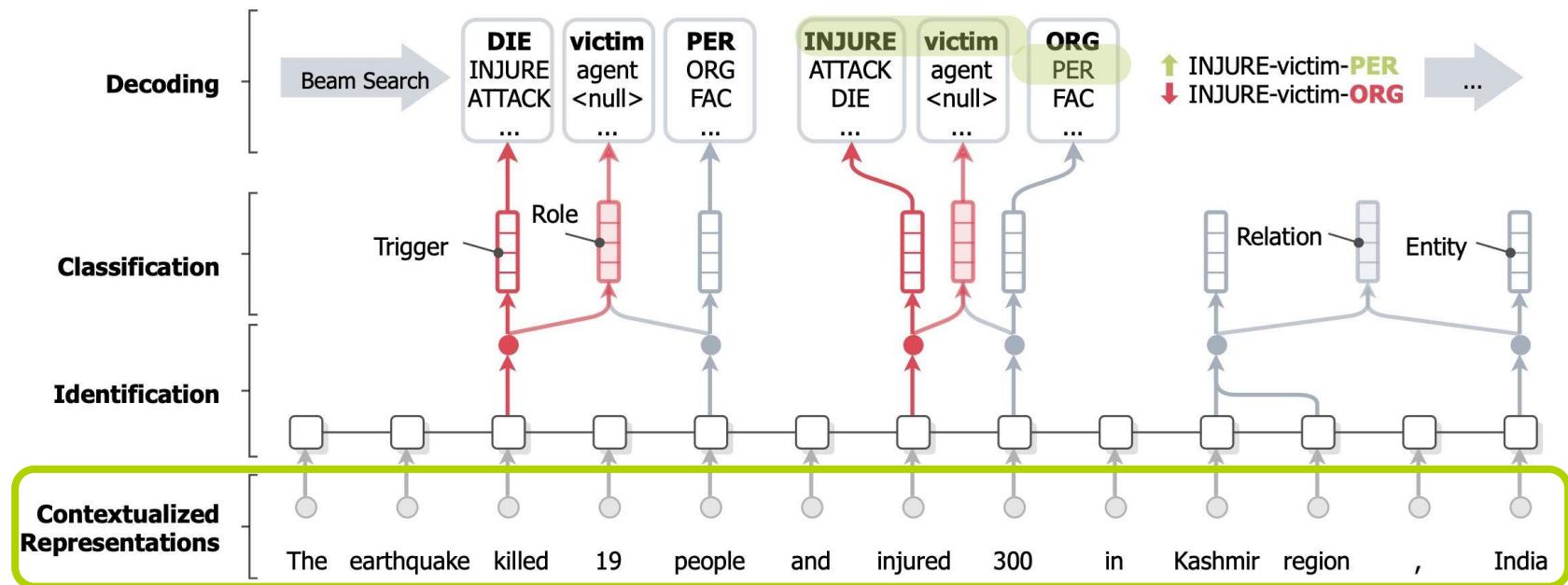
- Zhou et al. (2020): Temporal Information Extraction from Patterns
 - Step 1: Extract distant signals of contextualized events and their duration, frequency etc. via linguistic patterns
 - Step 2: further pre-train a language model with extracted instances

OneIE: An End-to-end Neural Model for IE



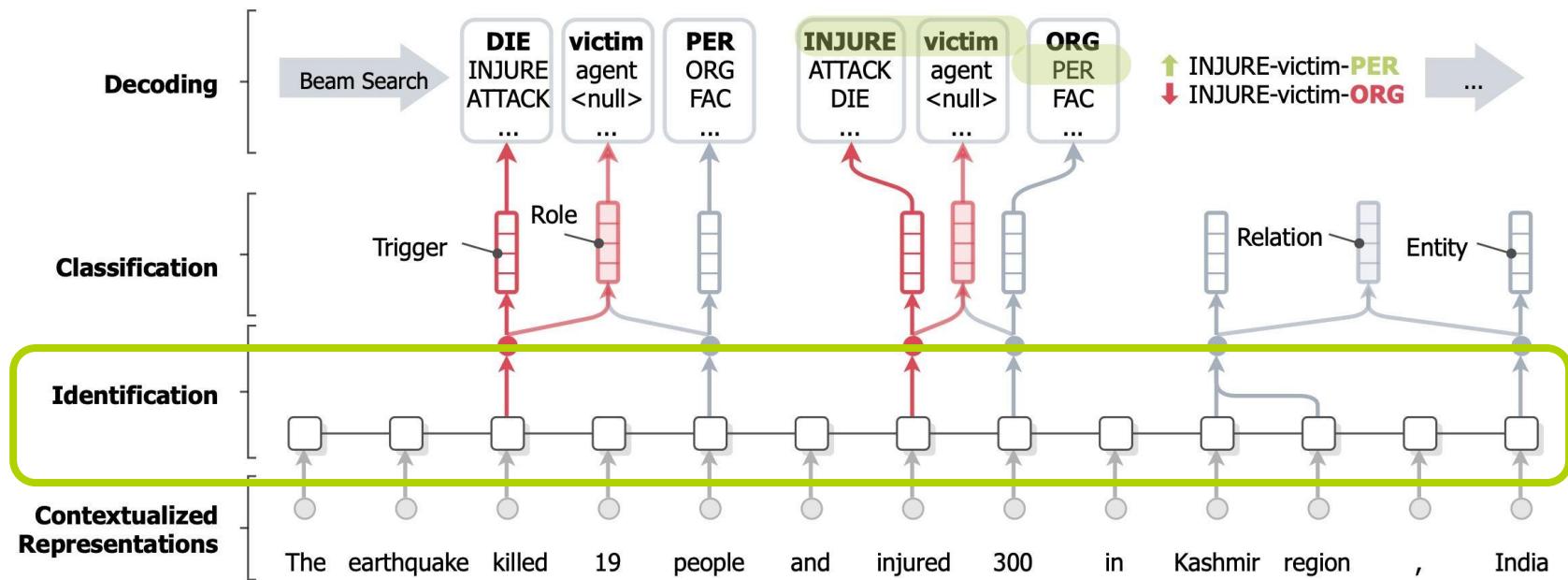
- Our OneIE framework extracts the information graph from a given sentence in four steps: encoding, identification, classification, and decoding

OneIE: An End-to-end Neural Model for IE



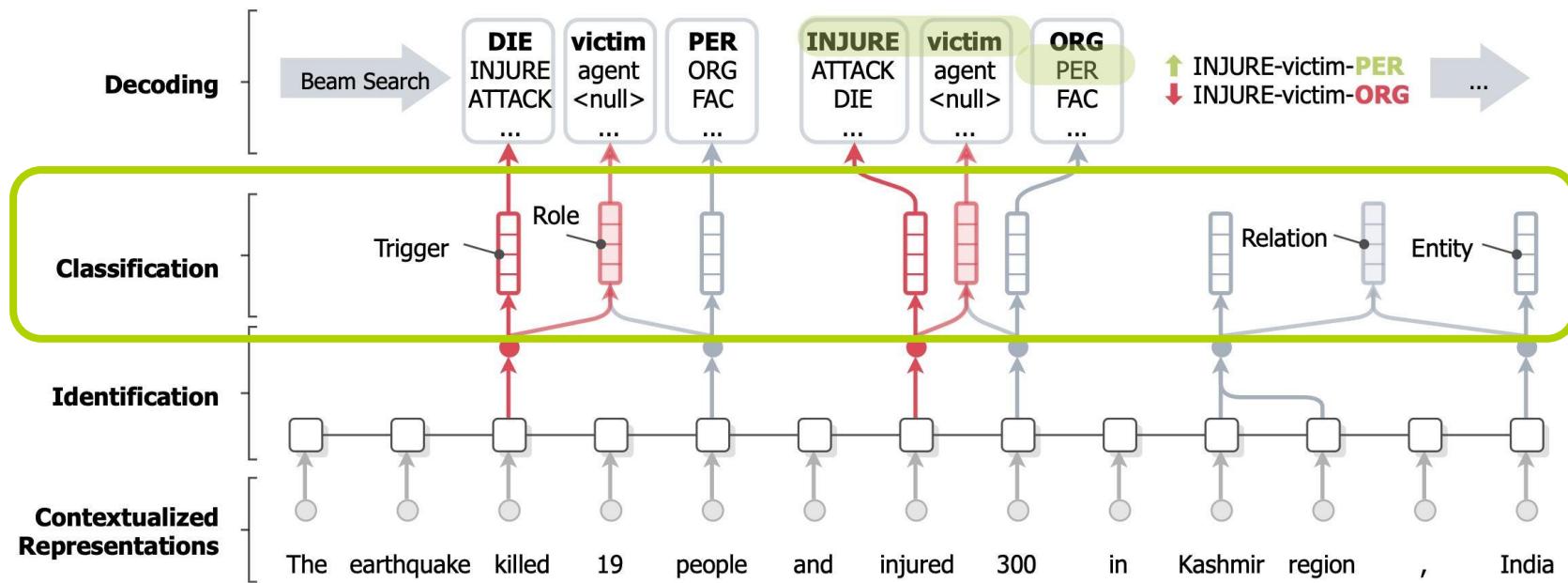
- **Encoding:** We use a BERT encoder to obtain the contextualized embedding of each token

OneIE: An End-to-end Neural Model for IE



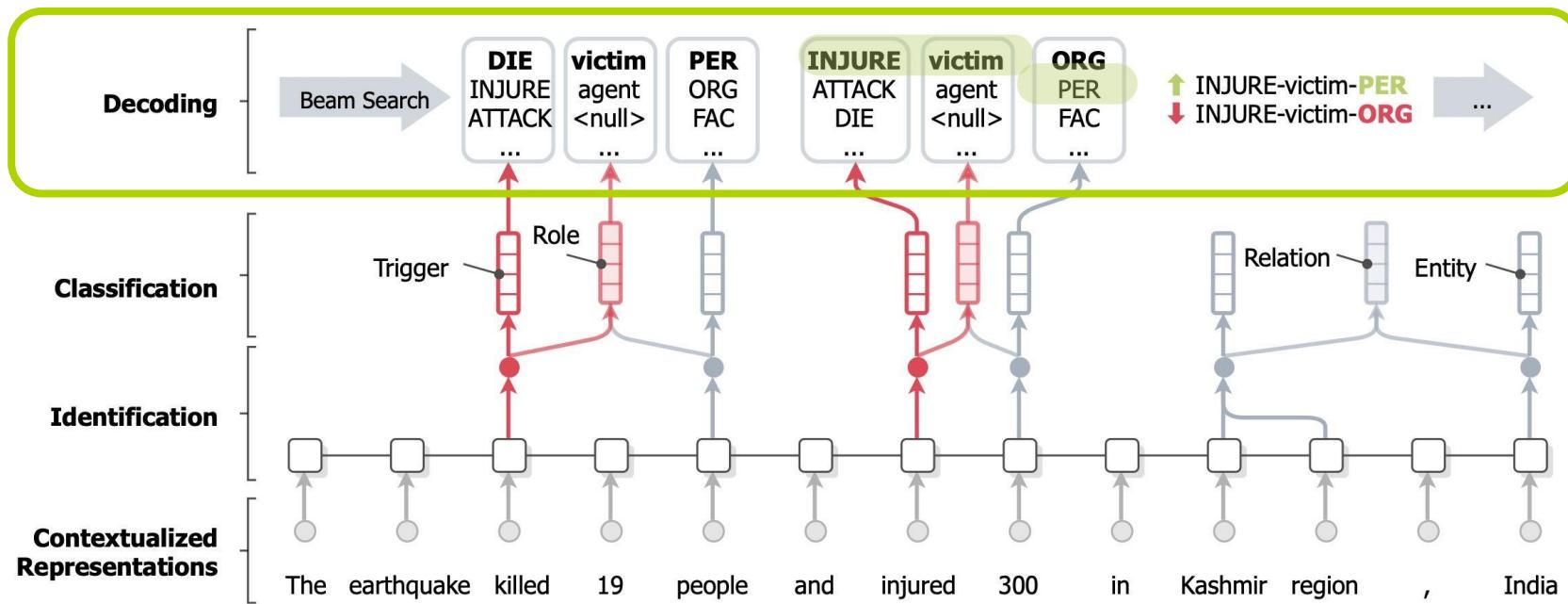
- **Identification:** We use CRF taggers to identify entity mentions and event triggers
- We define the identification loss as $\mathcal{L}^I = -\log p(\mathbf{z}|\mathbf{X})$

OneIE: An End-to-end Neural Model for IE



- **Classification:** We use task-specific feed-forward networks to calculate label scores for each node or edge
- We define the classification loss as $\mathcal{L}^t = -\frac{1}{N^t} \sum_{i=1}^{N^t} \mathbf{y}_i^t \log \hat{\mathbf{y}}_i^t$

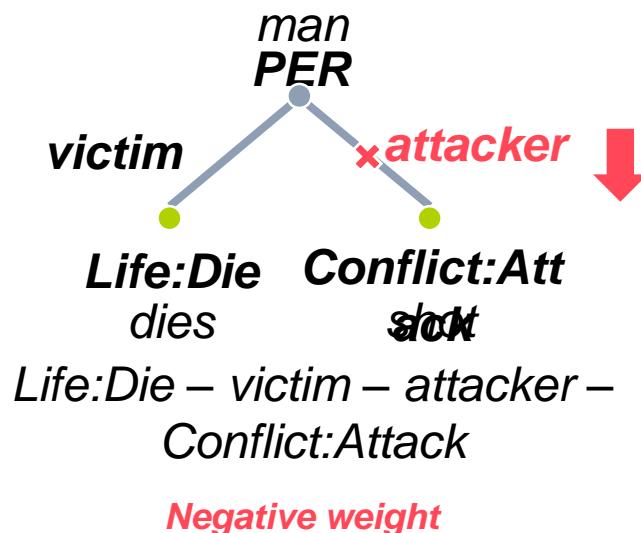
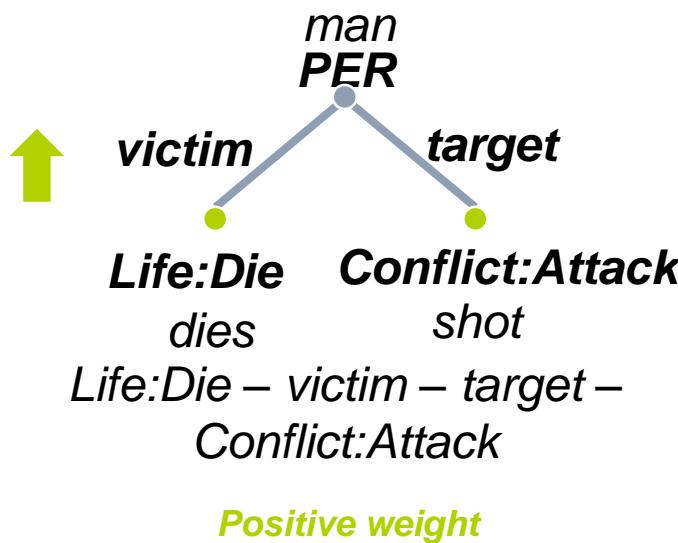
OneIE: An End-to-end Neural Model for IE



- **Decoding:** In the test phase, we use a beam search decoder to find the information graph with the **highest global score**

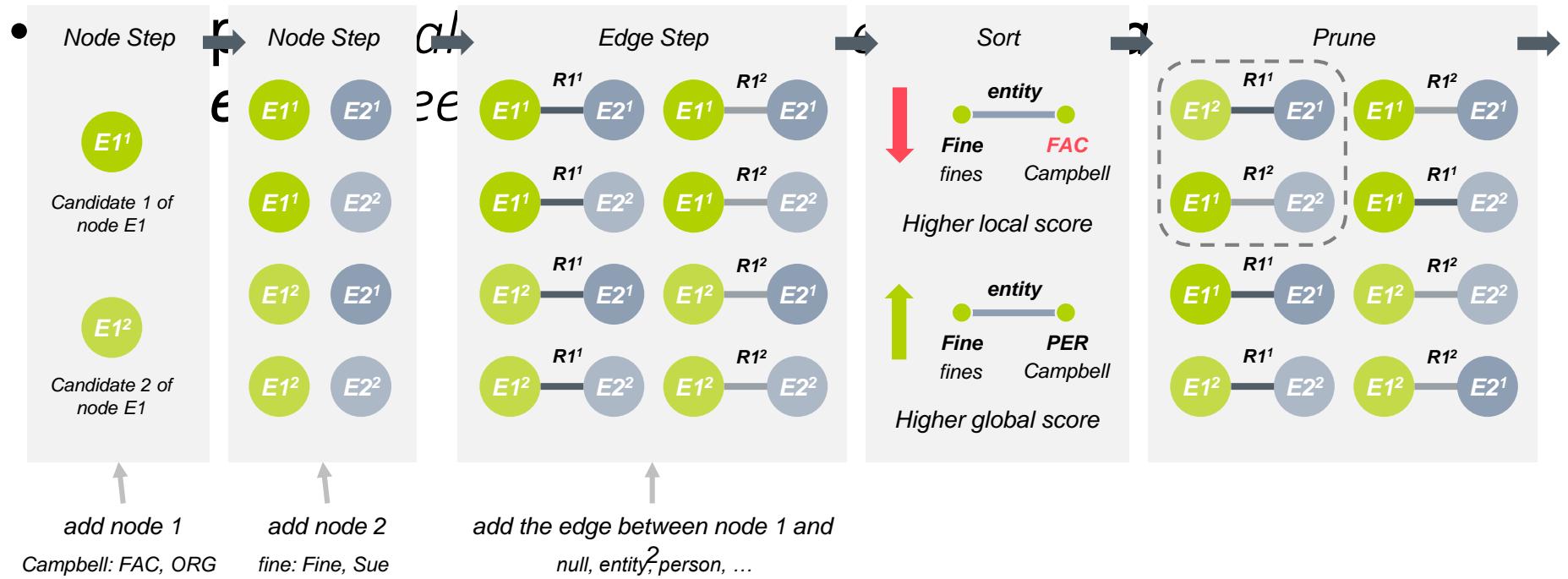
Incorporating Global Features

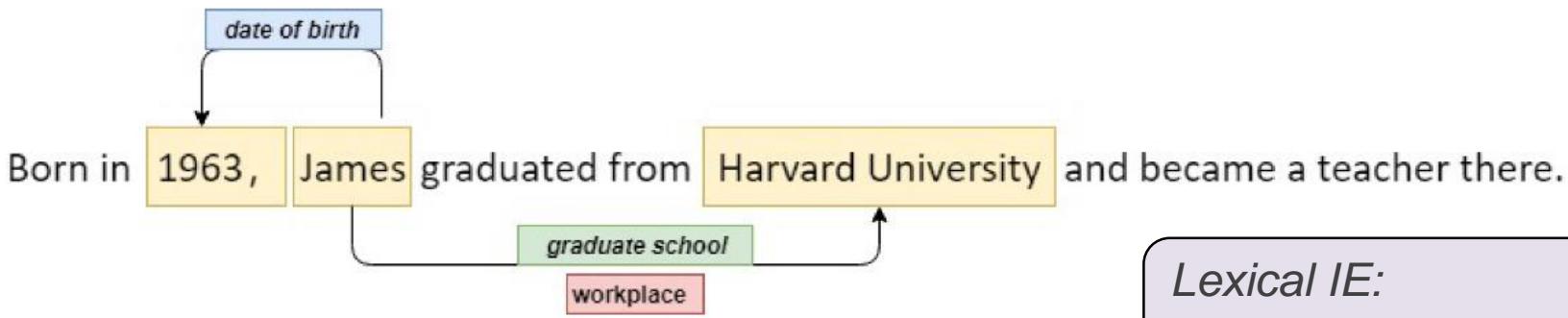
- We design a set of *global feature templates* (e.g., $\text{event_type}_1 - \text{role}_1 - \text{role}_2 - \text{event_type}_2$: an entity acts a role_1 argument for an event_type_1 event and a role_2 argument for an event_type_2 event in the same sentence)
- The model learns the *weight* of each feature during training



Decoding

- We use beam search to decode the information graph





- From the texts:
- 1 . Identify the concepts
 - Entities, events, terms, etc.
- 2. Identify the relations and other properties
 - Entity-entity / event-event
 - Temporal properties
 - etc.

Lexical IE:

- *Named entity recognition*
- *Entity/event typing*
- *Entity/event linking*

Relational IE:

- *Relation extraction*
 - *Entity / events*
 - *Sentence/Document*
 - *Temporal*
- *Coreference Resolution*

- Slide sources

- Most of the slides today came from a lecture of Dan Jurafsky's in Chris Manning and Dan Jurafsky's online NLP course at Stanford (covers very broad range of NLP and Machine Learning topics)

Last words

- As discussed in Sarawagi, traditional IE and web-based IE differ
 - Traditional IE: find relation between entities in one text (think of CMU Seminars for instance)
 - Web IE: find relation between "real-world" entities. Relations may occur on many different pages expressed in different ways
 - There are also tasks that are in between these two extremes
- Event extraction is like relation extraction
 - The difference is that we fill out templates
 - We have seen examples of these templates several times (for instance: outbreak – location – date)
 - Due to time, I am skipping the details of event extraction
 - In any case, how it is done is highly specific to the individual task to be performed

References

- <https://www.youtube.com/watch?v=z6twqnxF8Y8>
- <https://www.youtube.com/watch?v=3HNhhqj0di0>
- <https://www.youtube.com/watch?v=GoNlzi6JtHA>
- <https://www.youtube.com/watch?v=DnP5uN2EuWA>
- https://www.youtube.com/watch?v=8HL-Ap5_Axo
- <https://www.youtube.com/watch?v=25u7Zmczdl8>
- <https://www.youtube.com/watch?v=re5Aw6D7RNo>
- https://www.youtube.com/watch?v=PImNvfyVy_4
- <https://towardsdatascience.com/nlp-deep-learning-for-relation-extraction-9c5d13110afa>

References

<https://spacy.io/usage/v3>

<https://github.com/davidsbatista/Annotated-Semantic-Relationships-Datasets>

<https://www.youtube.com/watch?v=qpa99XSHak>

<https://slideplayer.com/slide/7234973/>

https://nlpprogress.com/english/relationship_extraction.html

<https://slideplayer.com/slide/5802639/>

<https://www.youtube.com/watch?v=18CTdWcJGL0>

<https://web.stanford.edu/~jurafsky/slp3/>

<https://slideplayer.com/slide/3367997/>

<https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html>

<https://aclanthology.org/2021.acl-long.344.mp4>



BITS Pilani
Pilani Campus

Natural Language Processing Applications

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in





Session 14: Information Extraction-Event Extraction Date – 17th March 2024

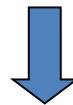
These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philipp Koehn**, Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

Agenda

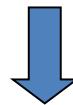
- Typical IE Pipeline
 - Temporal Information Extraction
 - What is Event
 - Event Extraction
 - Applications of Event Extraction
 - Event Process
 - Cross lingual Event Extraction
 - Case Study
-

A Typical IE Processing Pipeline

*Named Entity Recognition (NER) &
Shallow Parsing*



Reference Resolution



Relation Detection & Classification



Event Detection & Classification



Template Filling

Temporal Expression Extraction

- Temporal expressions are those that refer to absolute points in time, relative times, absolute durations, and sets of these.
- Absolute temporal expressions are those that can be mapped directly to calendar dates, times of day, or both.
- Relative temporal expressions map to particular times through some other reference point (as in a week from duration last Tuesday).
- Durations denote spans of time at varying levels of granularity (seconds, minutes, days, weeks, centuries, etc.).

Temporal Expression Extraction

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

Temporal Expression Extraction

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

Figure 18.19 Examples of temporal lexical triggers.

Temporal Expression Extraction

- The temporal expression recognition task consists of finding the start and end of all of the text spans that correspond to such temporal expressions.
- Rule-based approaches to temporal expression recognition use cascades of automata to recognize patterns at increasing levels of complexity.
- Tokens are first part-of-speech tagged, and then larger and larger chunks are recognized from the results from previous stages, based on patterns containing trigger words (e.g., February) or classes (e.g., MONTH).

Temporal Expression Extraction

```

# yesterday/today/tomorrow
$string =~ s/(((\$OT+the\$CT+\s+)?\$OT+day\$CT+\s+\$OT+(before|after)\$CT+\s+)?\$OT+\$TERelDayExpr\$CT+
(\s+\$OT+(morning|afternoon|evening|night)\$CT+)?)/<TIMEX$tever TYPE=\\"DATE\\">>\$1
</TIMEX$tever>/gio;

$string =~ s/(\$OT+\w+\$CT+\s+)<TIMEX$tever TYPE=\\"DATE\\">[^>]*>(\$OT+(Today|Tonight)\$CT+)
</TIMEX$tever>/\$1\$4/gso;

# this (morning/afternoon/evening)
$string =~ s/((\$OT+(early|late)\$CT+\s+)?\$OT+this\$CT+\s*\$OT+(morning|afternoon|evening)\$CT+)/
<TIMEX$tever TYPE=\\"DATE\\">>\$1</TIMEX$tever>/gosi;
$string =~ s/((\$OT+(early|late)\$CT+\s+)?\$OT+last\$CT+\s*\$OT+night\$CT+)/<TIMEX$tever
TYPE=\\"DATE\\">>\$1</TIMEX$tever>/gsio;

```

Figure 18.20 Perl fragment from the GUTime temporal tagging system in Tarsqi ([Verhagen et al., 2005](#)).

Temporal Expression Extraction

- Sequence-labeling approaches follow the same IOB scheme used for named entity tags, marking words that are either inside, outside or at the beginning of a TIMEX3-delimited temporal expression with the I, O, and B tags as follows:

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

O O O O B I O O O

Temporal Expression Extraction

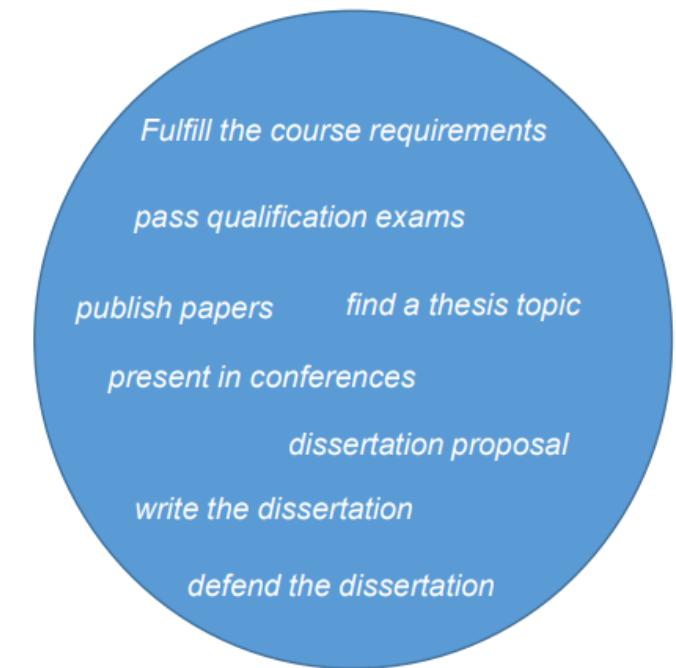
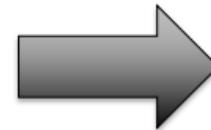
- Features are extracted from the token and its context, and a statistical sequence labeler is trained (any sequence model can be used).

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

Figure 18.21 Typical features used to train IOB-style temporal expression taggers.

What is Event?

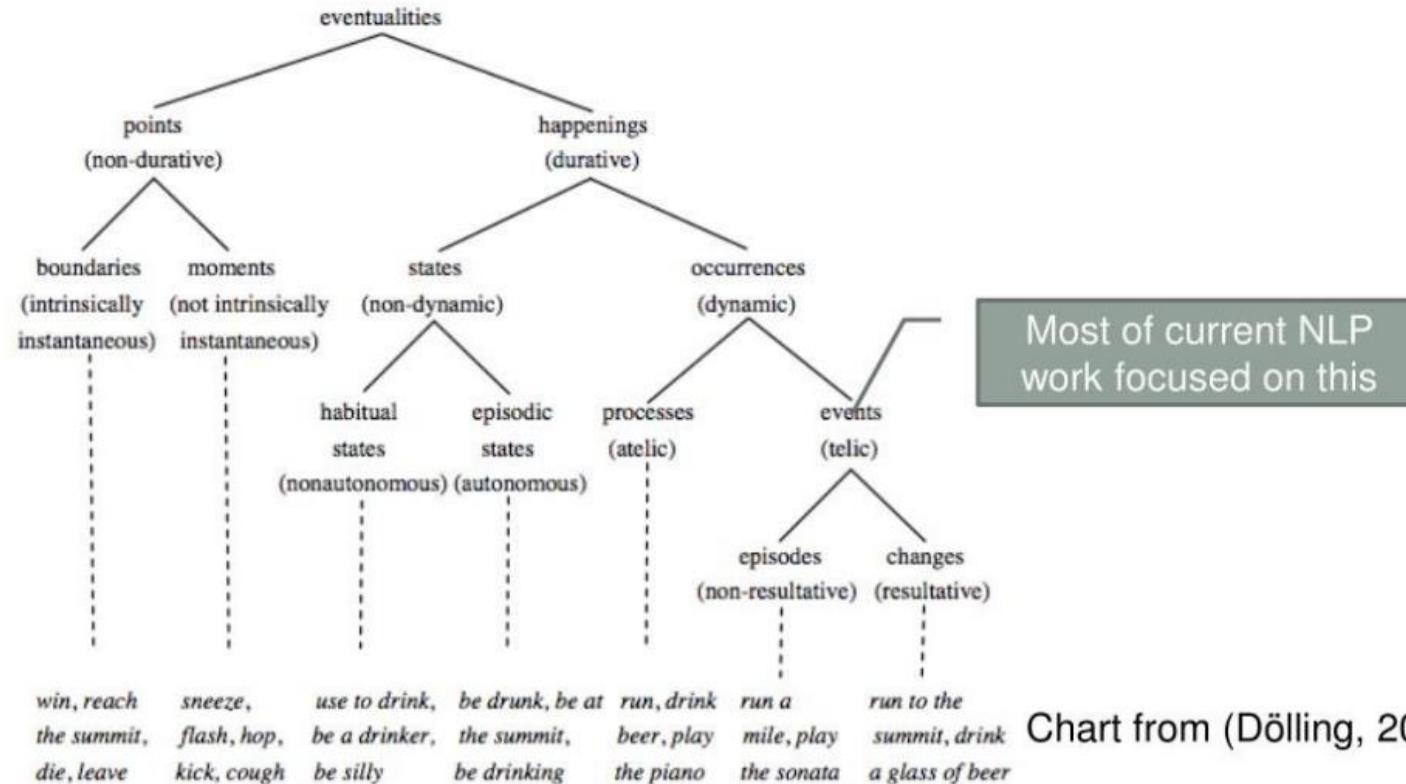
Earning a PhD in Computer Science typically takes around 5 years. It first involves **fulfilling the course requirements** and **passing qualification exams**. Then within several years, the student is expected to **find a thesis topic**, **publish several papers** about the topic and **present them in conferences**. The last one or two years are often about **completing the dissertation proposal**, **writing** and **defending the dissertation**.



Natural language understanding (NLU) has to deal with event understanding

What is Event?

- An Event is a specific occurrence involving participants.
- An Event is something that happens.
- An Event can frequently be described as a change of state.



What is Event?

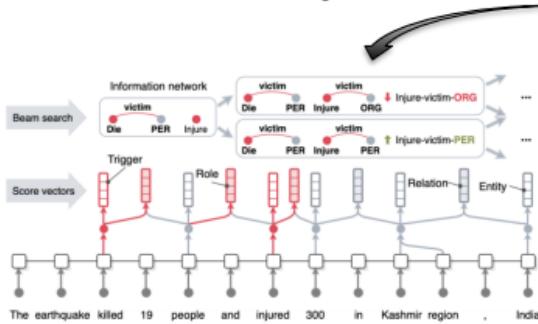
An **action** or a series of actions that happen at a specific location, within a period of time, and causes change(s) to the status of some object(s)

E.g.:

Jeff shaved my hair yesterday at home

How to recognize an event in text?

Supervised Methods



Bi-LSTM-CRF, Seq2Struct, etc.

- Annotated documents
- E.g, ACE-05, RED, ERE, etc...

Unsupervised Methods



Semantic Role Labeling
(Verb SRL / Nom SRL)

What is Event Extraction?

- Event extraction is to identify mentions of events in texts.
- An event mention is any expression denoting an event or state that can be assigned to a particular point, or interval, in time

NER versus Event Extraction

- Named Entity Recognition = identifying types of entities
- Event Extraction = identifying role relationships associated with events.

Paul Nelson killed John Smith.

Paul Nelson was killed by John Smith.

IBM purchased Microsoft.

IBM was purchased by Microsoft.

IBM was purchased on Tuesday by Microsoft

Relation vs Event Extraction

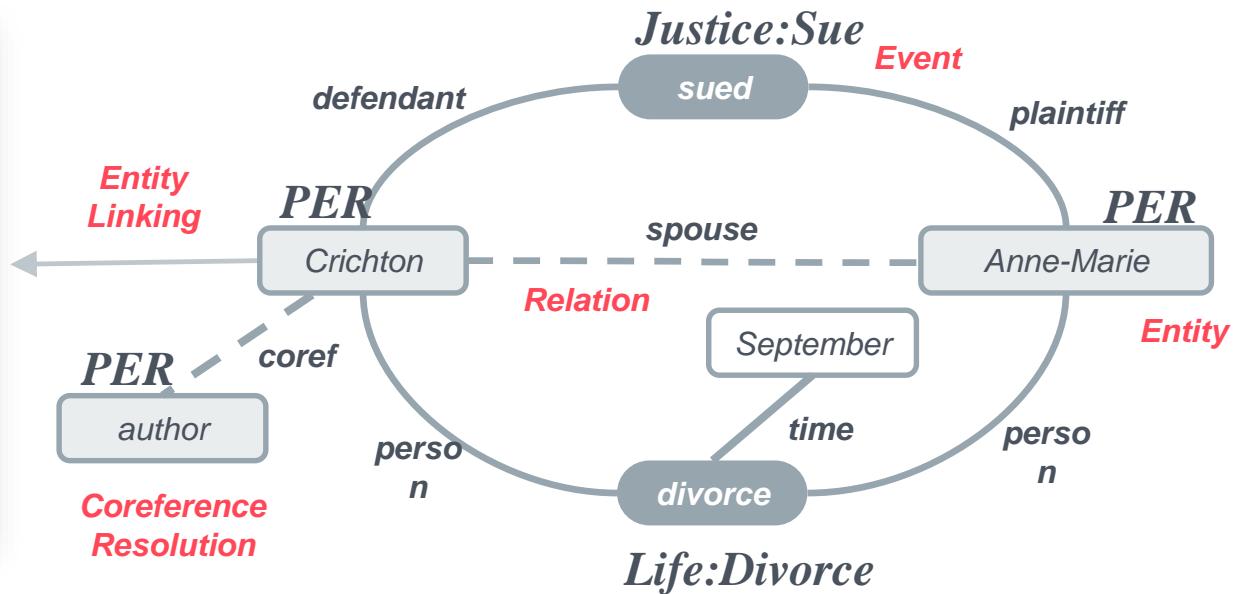
- Relationship Extraction identifies semantic relationships between two entities (for example, an affiliation relationship between a person and organization or a spousal relationship between two people)
- Event Extraction analyzes text for “*Who did What to Whom and Where and When,*” and finds events with one or more participants in each event.
- It also extracts the location and time of the event if the document contains them.
- Event Extraction finds additional information in text that provides a richer picture of people, organizations, places, and other entities beyond what Relationship extraction

Information Extraction Example



“

Anne-Marie sued Crichton, best known as the author of Jurassic Park, for divorce in September.

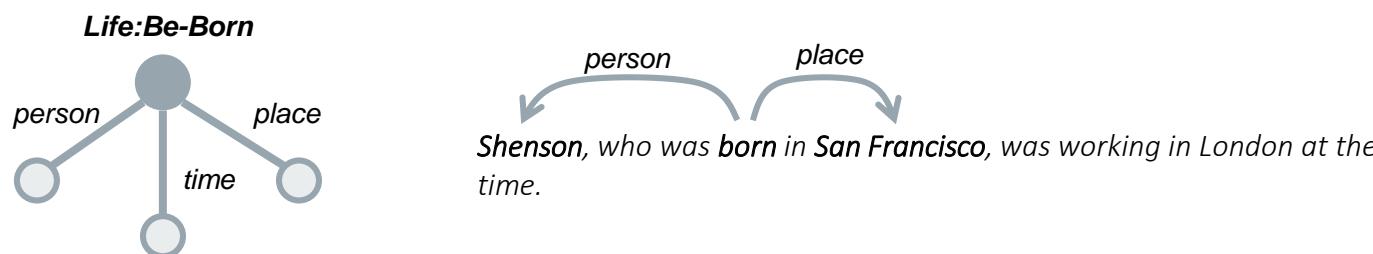


Information Extraction

Subtasks

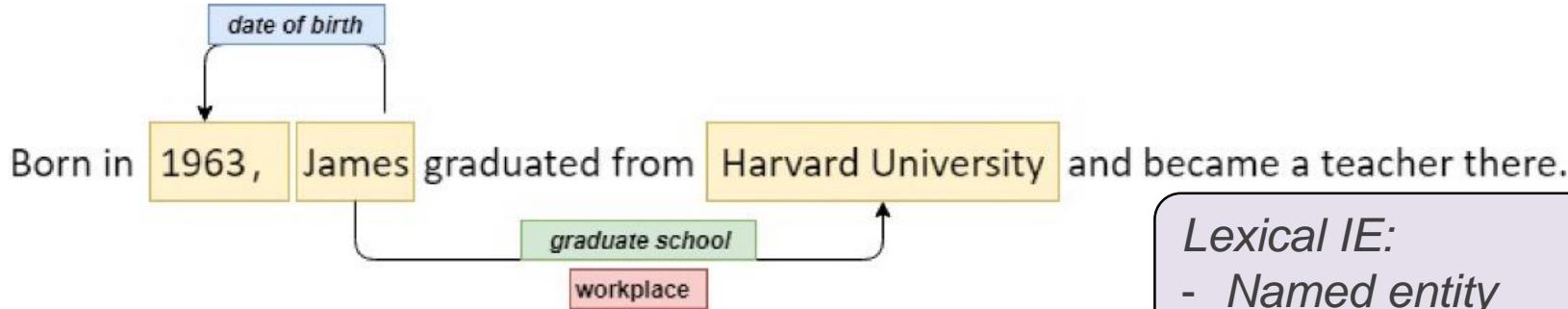


- **Entity Extraction** aims to identify entity mentions in text and classify them into pre-defined entity types.
- **Relation Extraction** is the task of assigning a relation type to an ordered pair of entity mentions.
- **Event Extraction** entails identifying and classifying event triggers and their arguments
 - Event triggers: the words or phrase that most clearly express event occurrences
 - Arguments: the words or phrases for participants in those events



- **Entity Coreference Resolution** is the task of resolving all entity mentions that refer to the same entity.
- **Event Coreference Resolution** is the task of resolving all event mentions that refer to the same event.

Information Extraction Subtasks



- From the texts:
- 1 . Identify the concepts
 - Entities, events, terms, etc.
- 2. Identify the relations and other properties
 - Entity-entity / event-event
 - Temporal properties
 - etc.

Lexical IE:

- *Named entity recognition*
- *Entity/event typing*
- *Entity/event linking*

Relational IE:

- *Relation extraction*
 - *Entity / events*
 - *Sentence/Document*
 - *Temporal*
- *Coreference Resolution*

NLU Applications of Event Extraction

Narrative prediction

One day Wesley's auntie came over to visit. He was happy to see her, because he liked to play with her. When she started to give his little sister attention, he got **jealous**. He got **angry** at his auntie and **bit** his sister's hand when she wasn't looking.

Then what might happen?

O1: He was **scolded**.



O2: She **gave him a cookie** for being so nice.



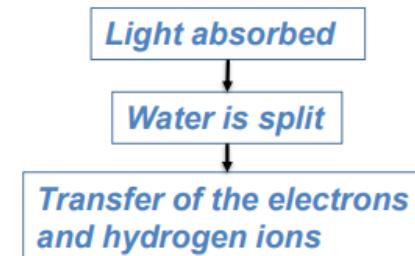
Machine comprehension

Water is split, providing a source of electrons and protons (hydrogen ions, H⁺) and giving off O₂ as a by-product. **Light absorbed** by chlorophyll drives a **transfer of the electrons and hydrogen ions** from water to an acceptor called NADP⁺.

What can the splitting of water lead to?

A: Light absorption

B: Transfer of ions



Input Text

Unstructured text depends 100% on language understanding.
Semi-structured text has some structure (layout) that can aid in understanding.

Unstructured Text

Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled “Embedded Commitment,” on Thursday, May 4th from 4-5:30 in PH 223D.

Semi-Structured Text

Laura Petitte
Department of Psychology
McGill University

Thursday, May 4, 1995
12:00 pm
Baker Hall 355

Name: Dr. Jeffrey D. Hermes

Affiliation: Department of AutoImmune Diseases

Research & Biophysical Chemistry Merch Research Laboratories

Title: "MHC Class II: A Target for Specific
Immunomodulation of the Immune Response"

Host/e-mail: Robert Murphy

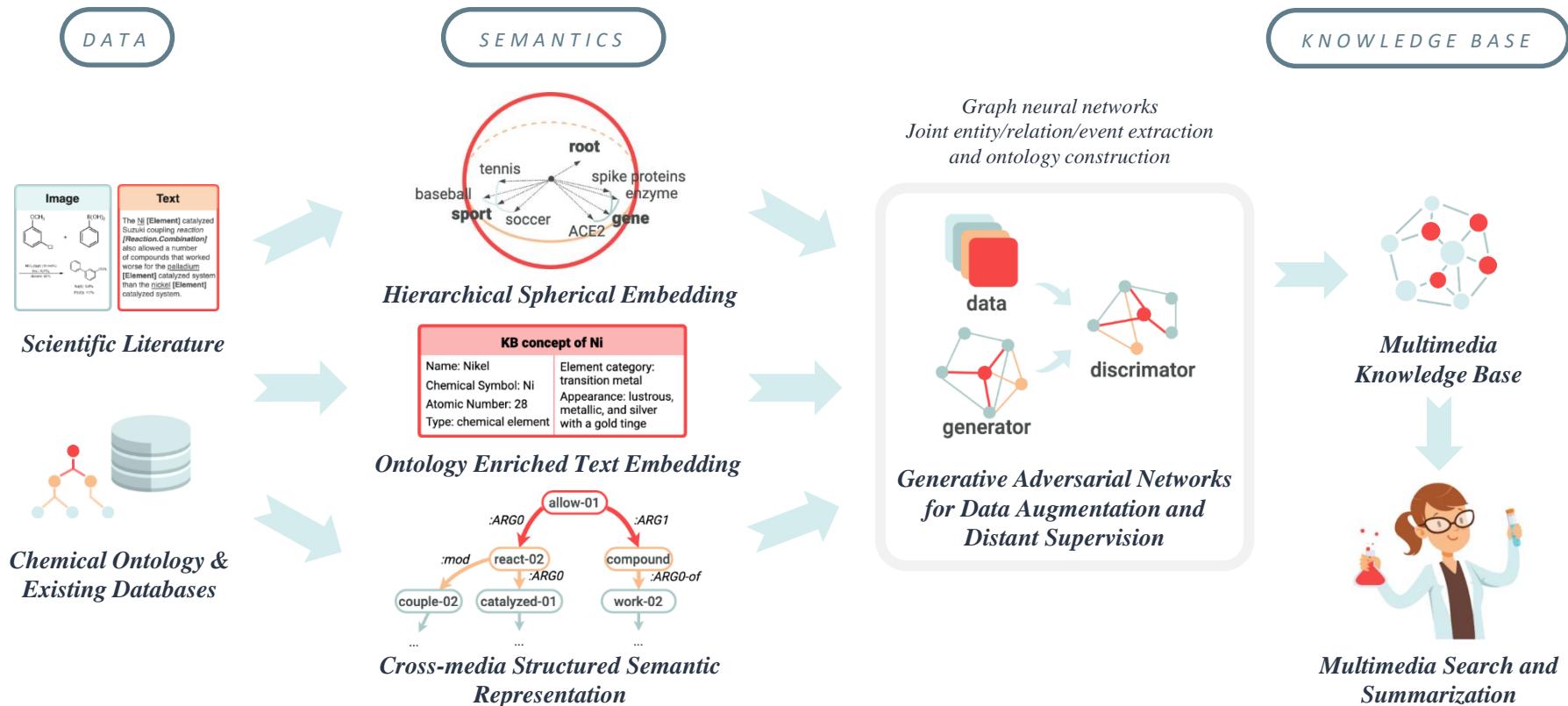
Date: Wednesday, May 3, 1995

Time: 3:30 p.m.

Place: Mellon Institute Conference Room

Sponsor: MERCK RESEARCH LABORATORIES

GOAL: converting unstructured DATA to structured KNOWLEDGE

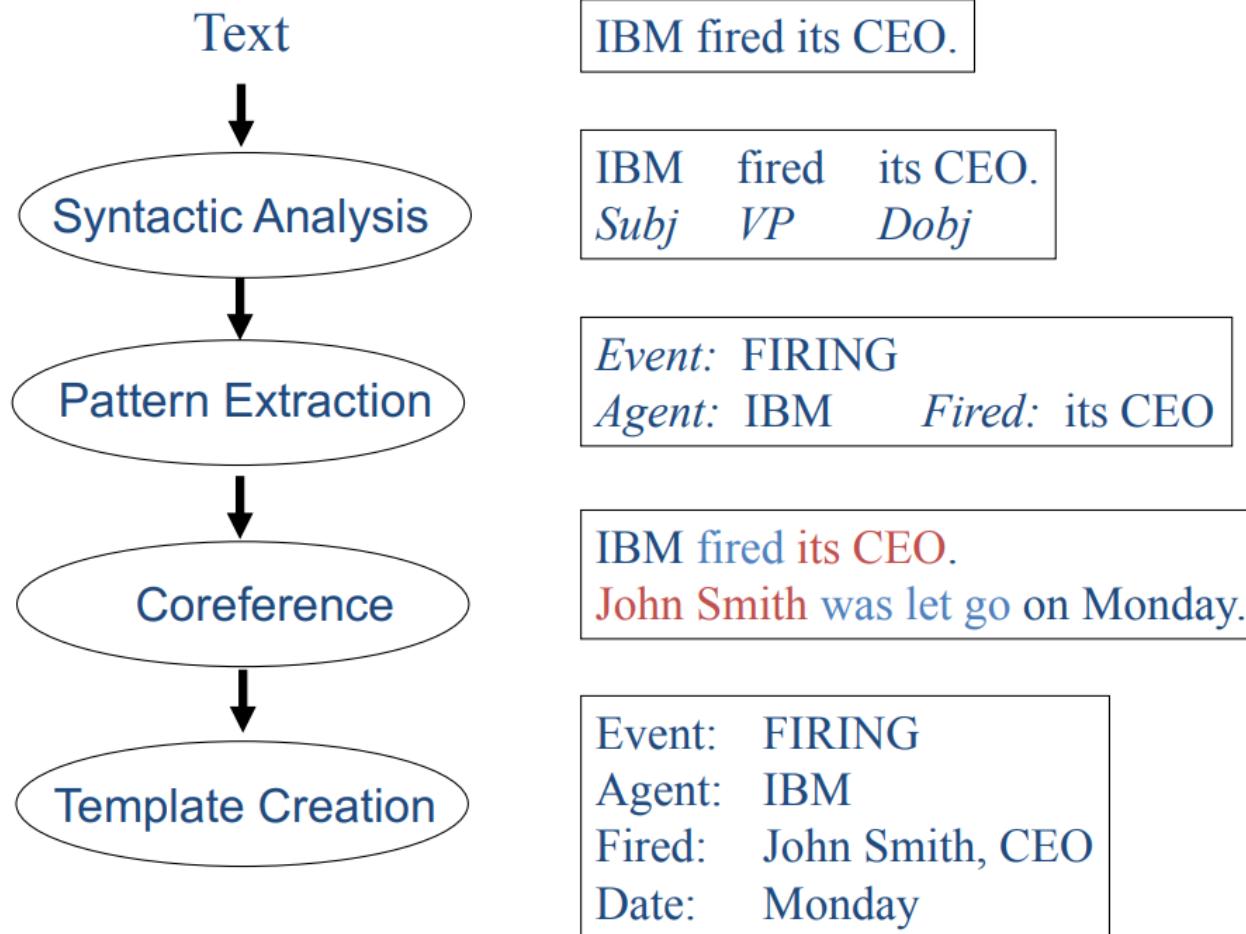


Event mention Extraction

- An event is specific occurrence that implies a change of states
- event trigger:** the main word which most clearly expresses an event occurrence
- event arguments:** the mentions that are involved in an event (participants)
- event mention:** a phrase or sentence within which an event is described, including trigger and arguments
- Automatic Content Extraction defined 8 types of events, with 33 subtypes

ACE event type/subtype	Argument, role=victim	trigger	Event Mention Example
Life/Die		Kurt Schork	died in Sierra Leone yesterday
Transaction/Transfer		GM	sold the company in Nov 1998 to LLC
Movement/Transport		Homeless people	have been moved to schools
Business/Start-Org	Schweitzer	founded	a hospital in 1913
Conflict/Attack		the attack	on Gaza killed 13
Contact/Meet		Arafat's cabinet	met for 4 hours
Personnel/Start-Position		She later	recruited the nursing student
Justice/Arrest	Faison	was wrongly arrested	on suspicion of murder

Pattern-based Template-Filling Pipeline



Supervised Learning Approach

- Build a classifier as a sequence tagging model.
- Each document is processed sequentially and each token is labeled as Extraction or Non-Extraction.
- Ex: B (beginning), I (inside), or O (outside) tags.
- Features are usually simple: e.g., words, POS tags, orthography, and a small context window of preceding/following words

Supervised Event Mention Extraction: Methods

- Staged classifiers

- Trigger Classifier
 - to distinguish event instances from non-events, to classify event instances by type
- Argument Classifier
 - to distinguish arguments from non-arguments
- Role Classifier
 - to classify arguments by argument role
- Reportable-Event Classifier
 - to determine whether there is a reportable event instance
- Can choose any supervised learning methods such as MaxEnt and SVMs

(Ji and Grishman, 2008)

Typical Event Mention Extraction Features

- Trigger Labeling

- Lexical

- Tokens and POS tags of candidate trigger and context words

- Dictionaries

- Trigger list, synonym gazetteers

- Syntactic

- the depth of the trigger in the parse tree
 - the path from the node of the trigger to the root in the parse tree
 - the phrase structure expanded by the parent node of the trigger
 - the phrase type of the trigger

- Entity

- the entity type of the syntactically nearest entity to the trigger in the parse tree
 - the entity type of the physically nearest entity to the trigger in the sentence

- Argument Labeling

- Event type and trigger

- Trigger tokens
 - Event type and subtype

- Entity

- Entity type and subtype
 - Head word of the entity mention

- Context

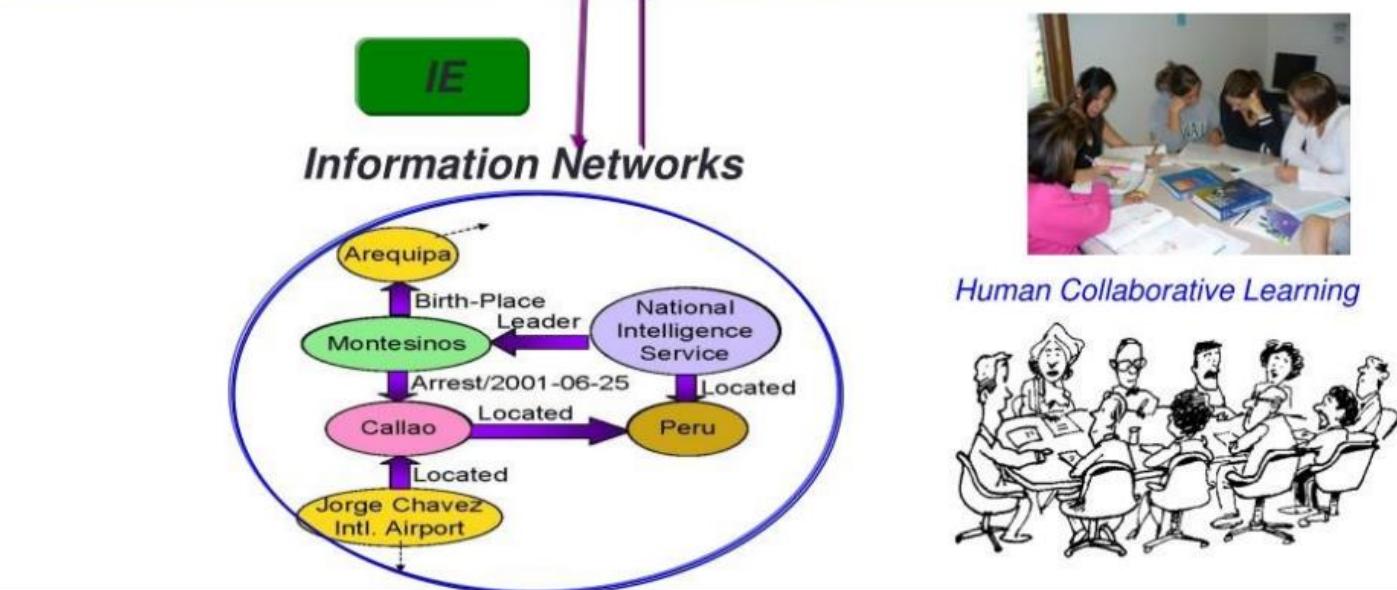
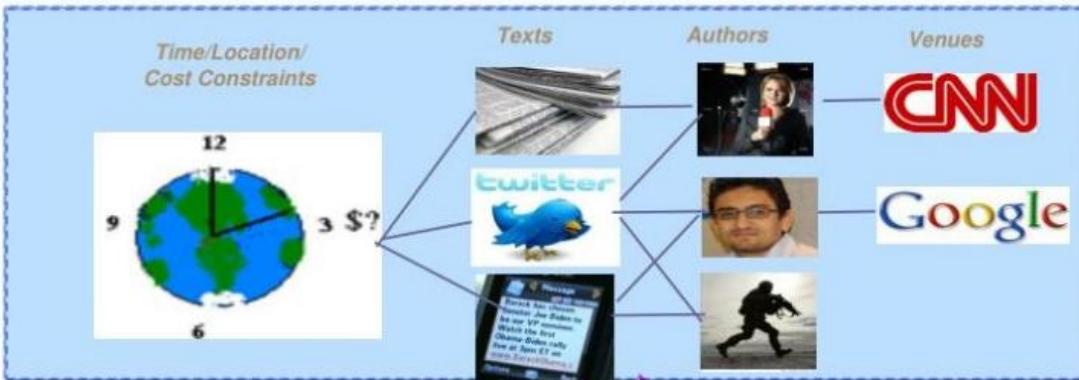
- Context words of the argument candidate

- Syntactic

- the phrase structure expanding the parent of the trigger
 - the relative position of the entity regarding to the trigger (before or after)
 - the minimal path from the entity to the trigger
 - the shortest length from the entity to the trigger in the parse tree

(Chen and Ji, 2009)

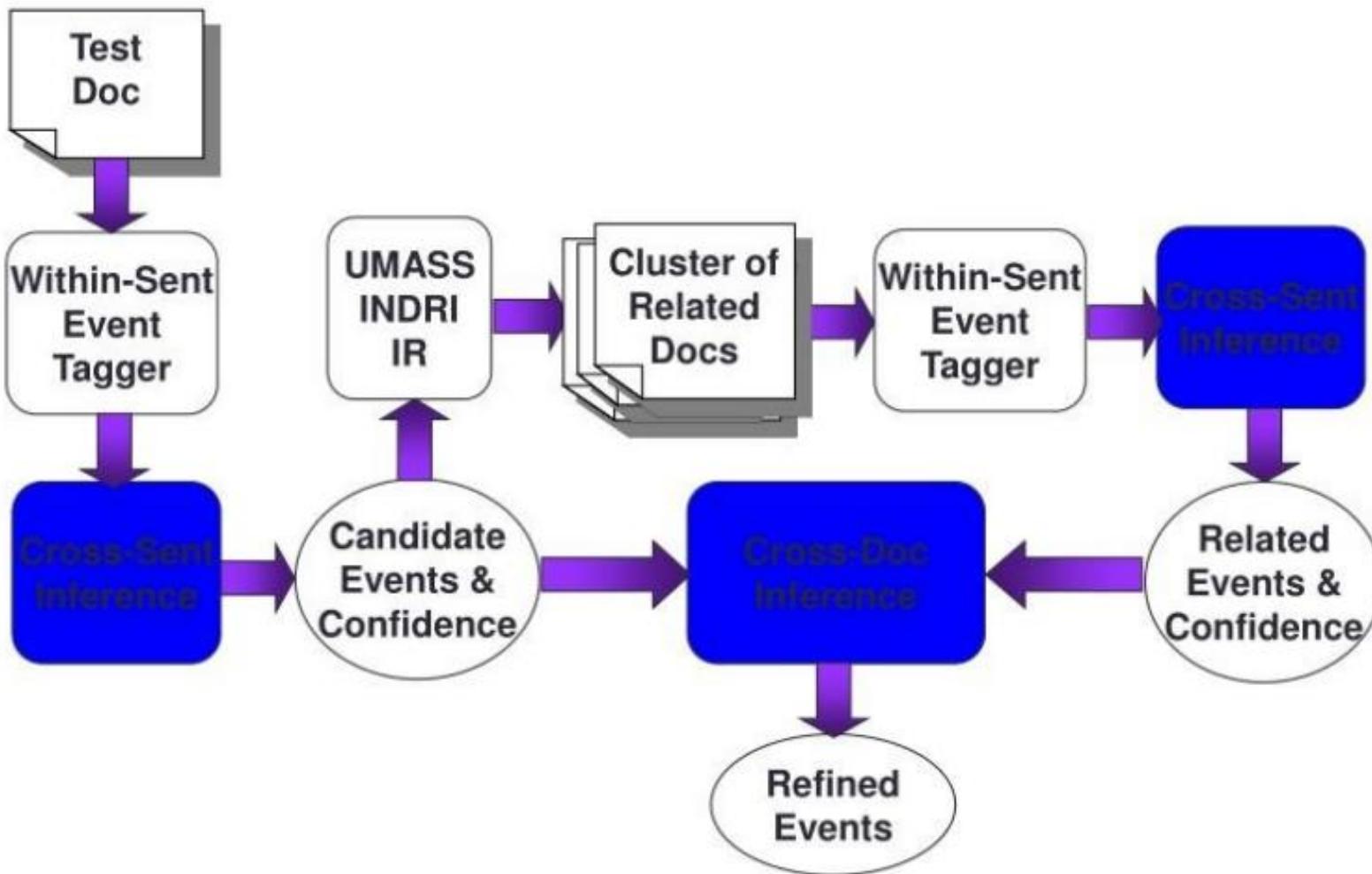
IE in Rich Contexts



Capture Information Redundancy

- When the data grows beyond some certain size, IE task is naturally embedded in rich contexts; the extracted facts become inter-dependent
- Leverage Information Redundancy from:
 - Large Scale Data (Chen and Ji, 2011)
 - Background Knowledge (Chan and Roth, 2010; Rahman and Ng, 2011)
 - Inter-connected facts (Li and Ji, 2011; Li et al., 2011; e.g. Roth and Yih, 2004; Gupta and Ji, 2009; Liao and Grishman, 2010; Hong et al., 2011)
 - Diverse Documents (Downey et al., 2005; Yangarber, 2006; Patwardhan and Riloff, 2009; Mann, 2007; Ji and Grishman, 2008)
 - Diverse Systems (Tamang and Ji, 2011)
 - Diverse Languages (Snover et al., 2011)
 - Diverse Data Modalities (text, image, speech, video...)
- But how? Such knowledge might be overwhelming...

Cross Sentences/Doc Event Inferencing



Within Sentence Extraction

1. Pattern matching

- Build a pattern from each ACE training example of an event
 - British and US forces reported gains in the advance on Baghdad
→ PER report gain in advance on LOC

2. MaxEnt models

① Trigger Classifier

- to distinguish event instances from non-events, to classify event instances by type

② Argument Classifier

- to distinguish arguments from non-arguments

③ Role Classifier

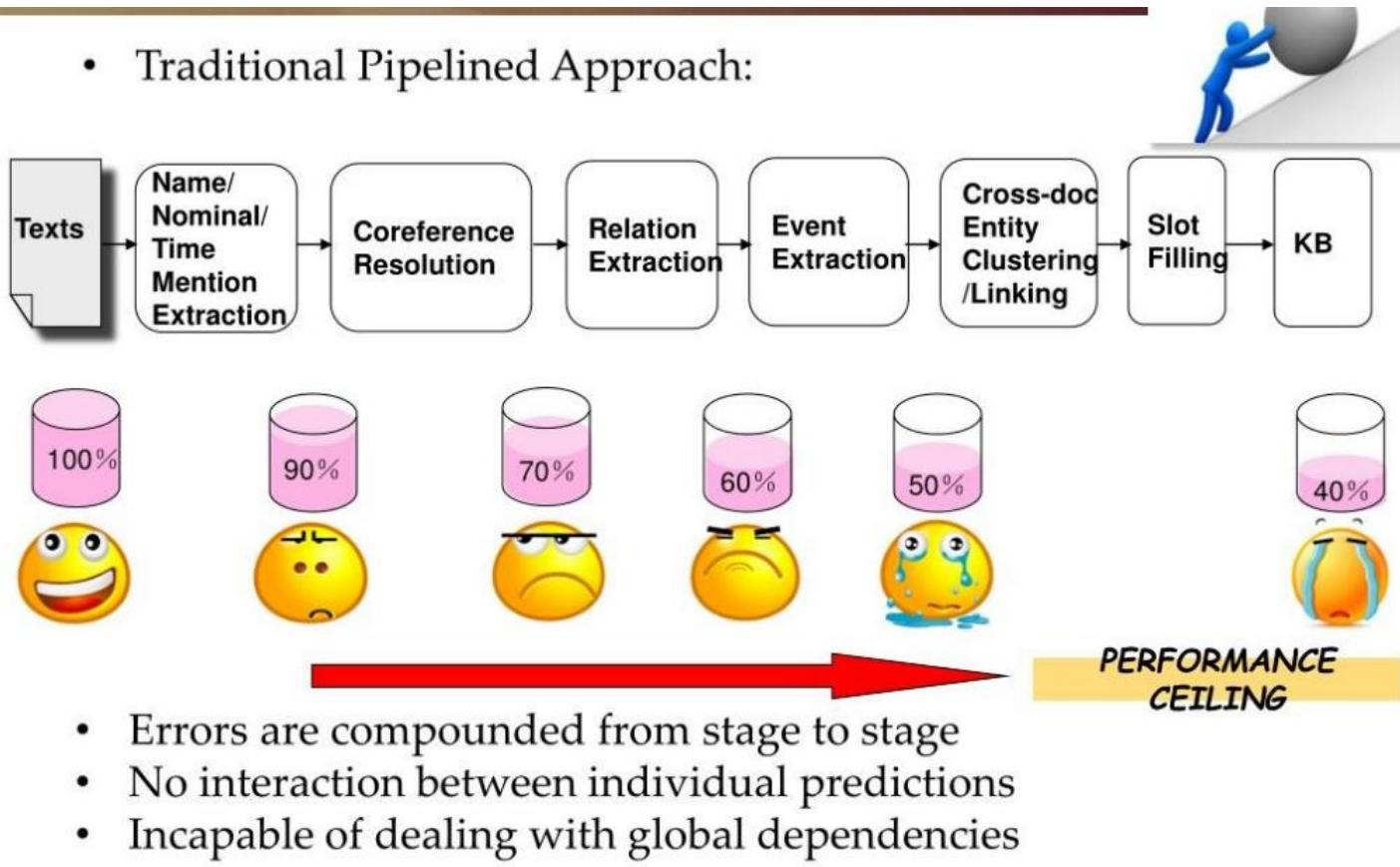
- to classify arguments by argument role

④ Reportable-Event Classifier

- to determine whether there is a reportable event instance

Event Mention Extraction

- Traditional Pipelined Approach:

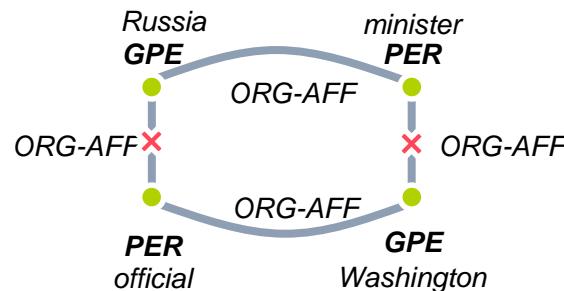


- Errors are compounded from stage to stage
- No interaction between individual predictions
- Incapable of dealing with global dependencies

Motivation

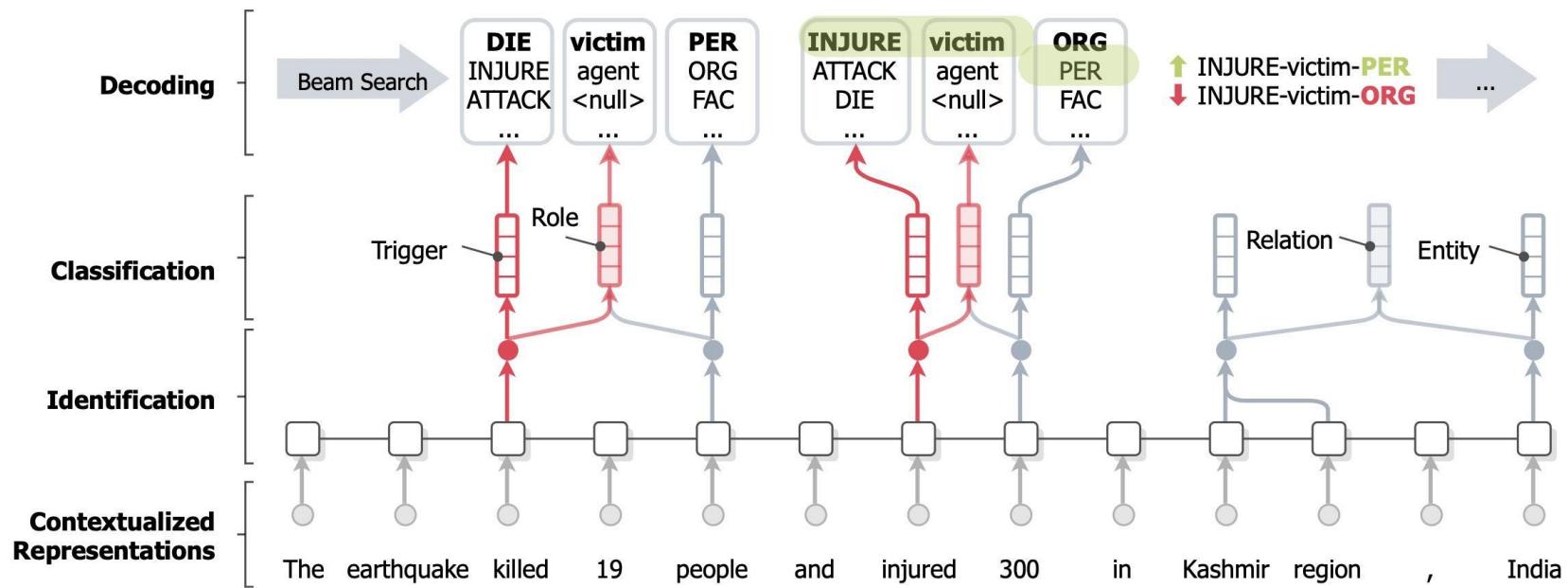
- Pipeline models suffer from the **error propagation problem** and disallow interactions among components.
- Existing neural models do not explicitly model **cross-subtask and cross-instance interactions** among knowledge elements.

*Russia's foreign **minister** expressed outrage at suggestions from a top **Washington official** last week...*



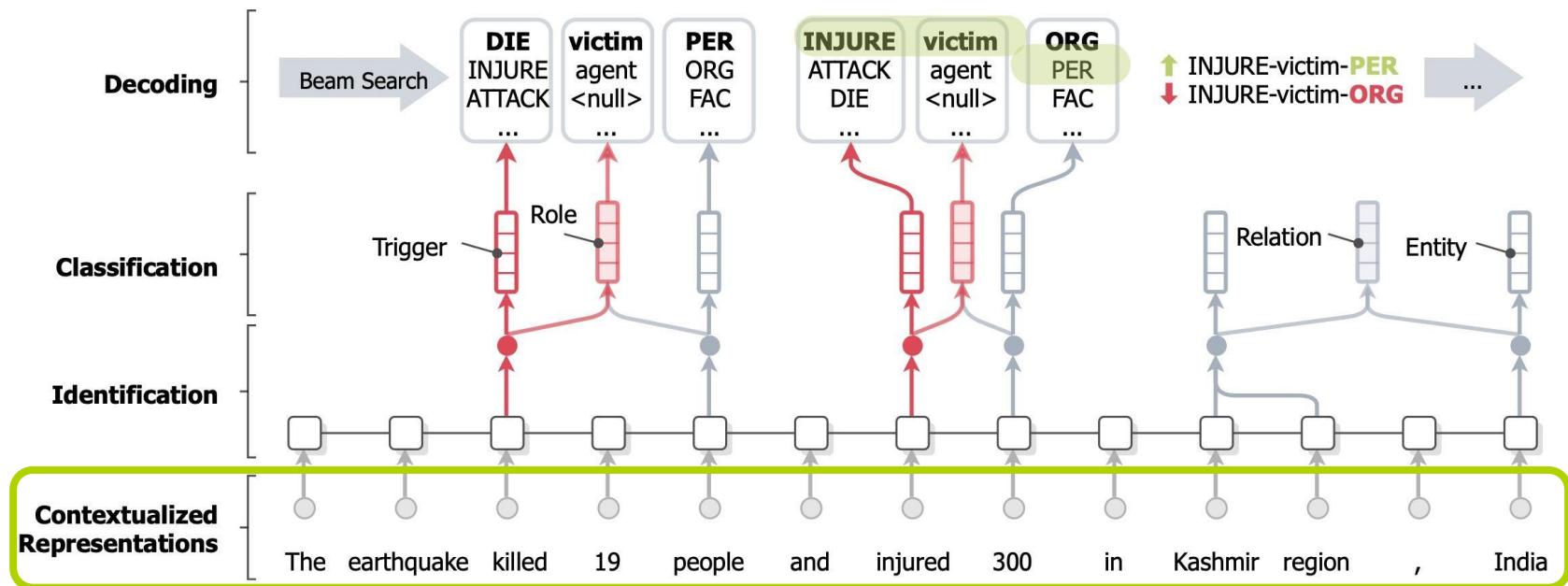
35

OneIE: An End-to-end Neural Model for IE



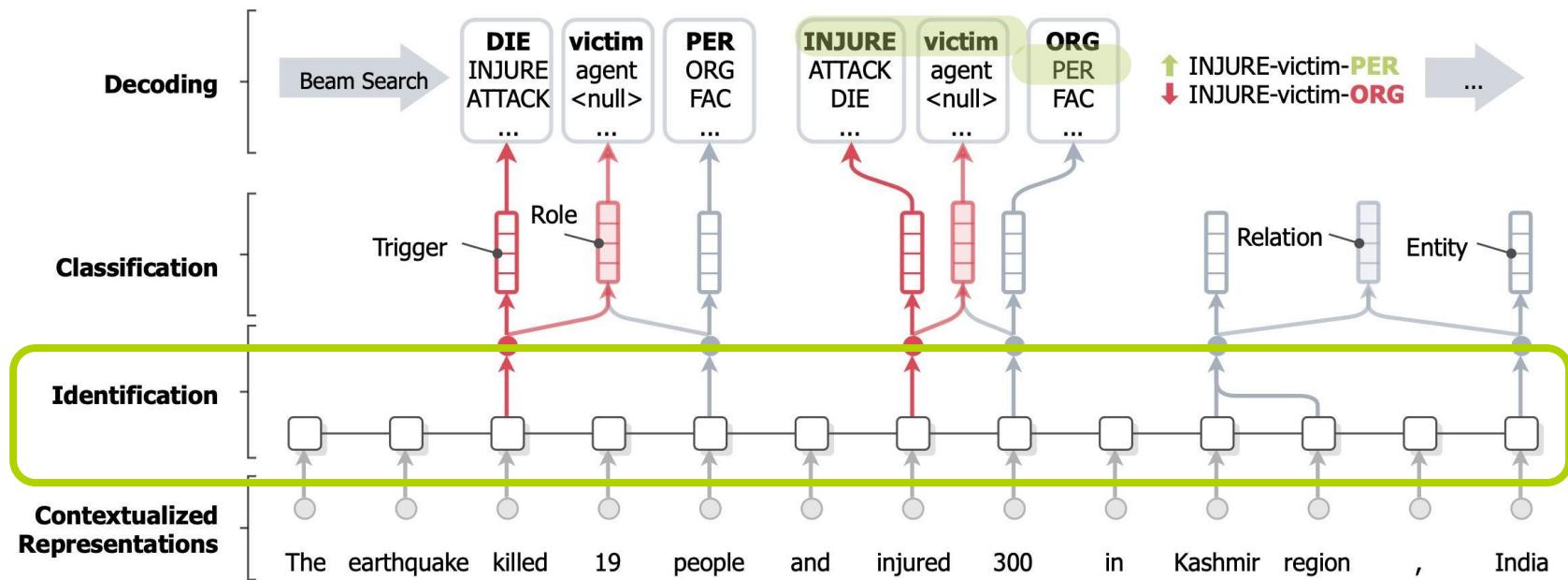
- Our OneIE framework extracts the information graph from a given sentence in four steps: encoding, identification, classification, and decoding

OneIE: An End-to-end Neural Model for IE



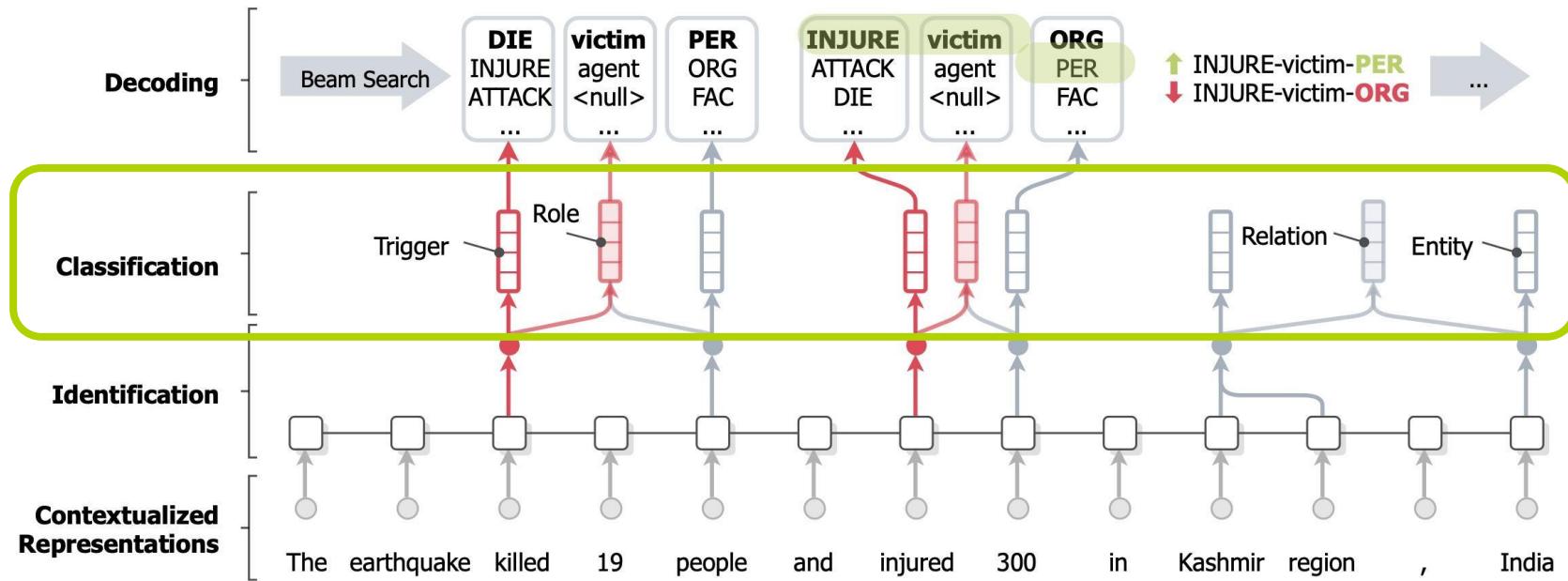
- **Encoding:** We use a BERT encoder to obtain the contextualized embedding of each token

OneIE: An End-to-end Neural Model for IE



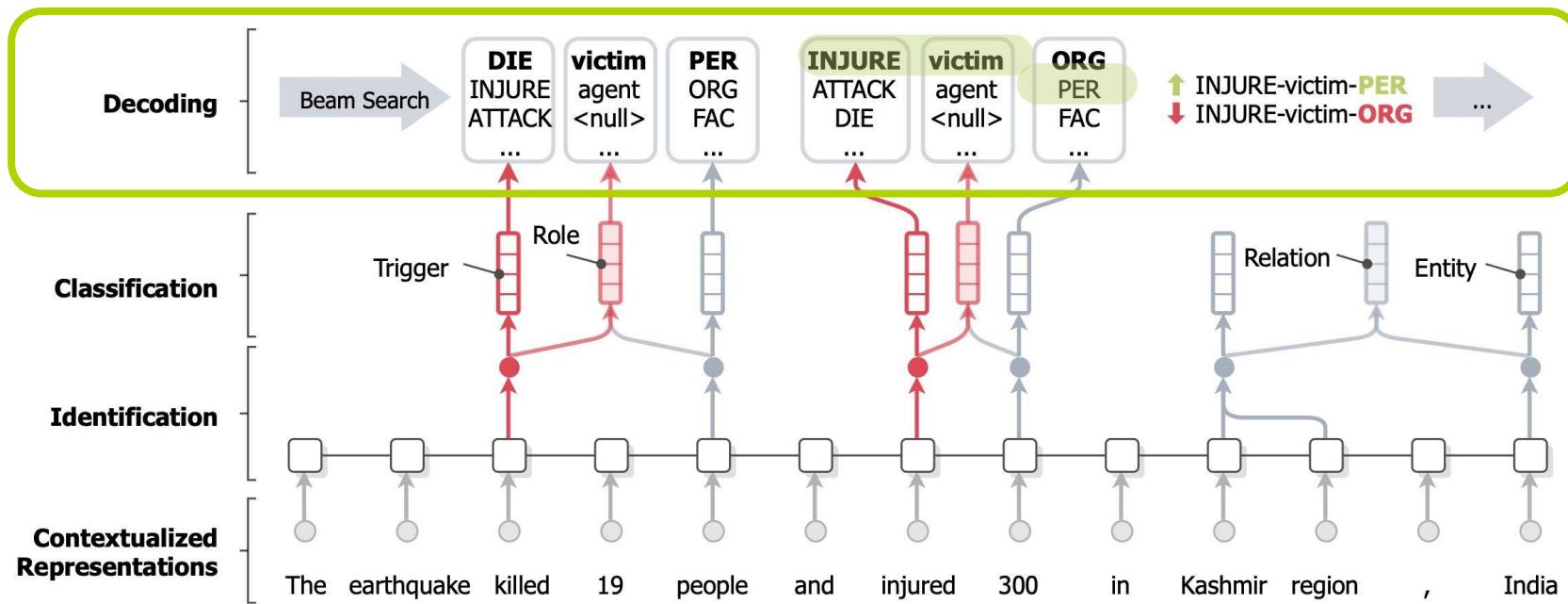
- **Identification:** We use CRF taggers to identify entity mentions and event triggers
- We define the identification loss as $\mathcal{L}^I = -\log p(\mathbf{z}|\mathbf{X})$

OneIE: An End-to-end Neural Model for IE



- **Classification:** We use task-specific feed-forward networks to calculate label scores for each node or edge
- We define the classification loss as $\mathcal{L}^t = -\frac{1}{N^t} \sum_{i=1}^{N^t} \mathbf{y}_i^t \log \hat{\mathbf{y}}_i^t$

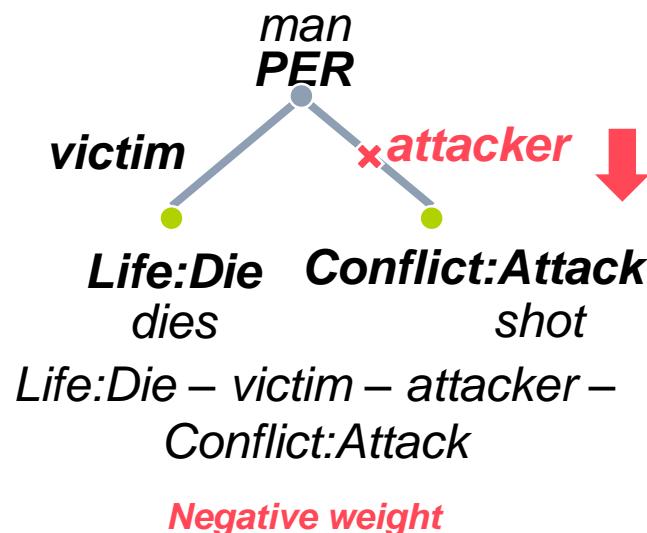
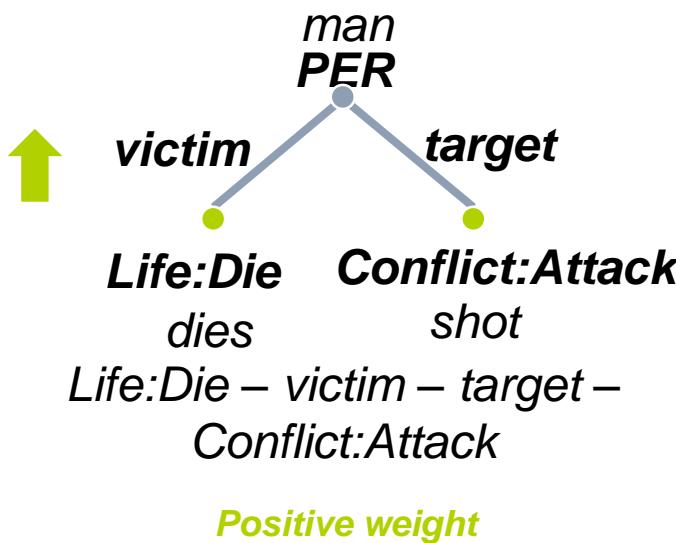
OneIE: An End-to-end Neural Model for IE



- **Decoding:** In the test phase, we use a beam search decoder to find the information graph with the **highest global score**

Incorporating Global Features

- We design a set of *global feature templates* (e.g., $\text{event_type}_1 - \text{role}_1 - \text{role}_2 - \text{event_type}_2$: an entity acts a role_1 argument for an event_type_1 event and a role_2 argument for an event_type_2 event in the same sentence)
- The model learns the *weight* of each feature during training



Salient Global Features

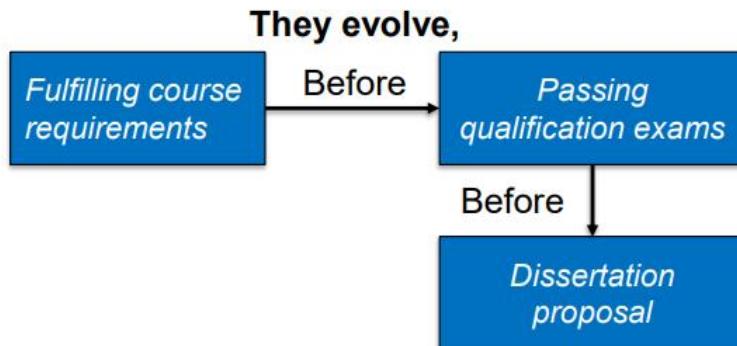
- Salient positive and negative global features learned by the model
- Global features are explainable

Features	Weight
1 A Transport event has only one Destination argument	2.61
2 An Attack event has only one Place argument	2.31
3 A PER-SOC relation exists between two PER entities	1.51
4 A Beneficiary argument is a PER entity	0.93
5 An entity has an ORG-AFF relation with multiple entities	-3.21
6 An event has two Place arguments	-2.47
7 A Transport event has multiple Destination argument	-2.25
8 An entity has a GEN-AFF relation with multiple entities	-2.02

Events Process

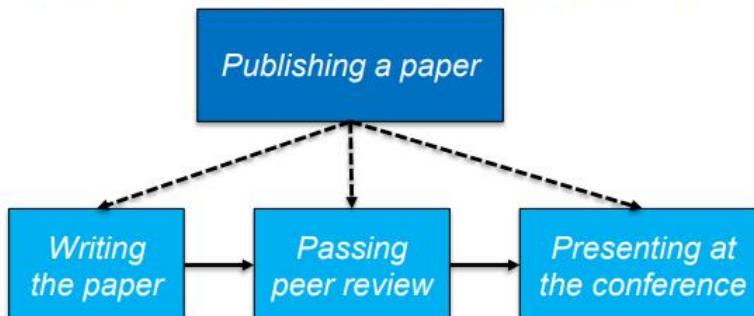
Extraction only is not enough.

Events are **NOT simple, static predicates.**



and are always directed by specific intents or central goals [Zacks et al. *Nature Neuroscience*, 2001]

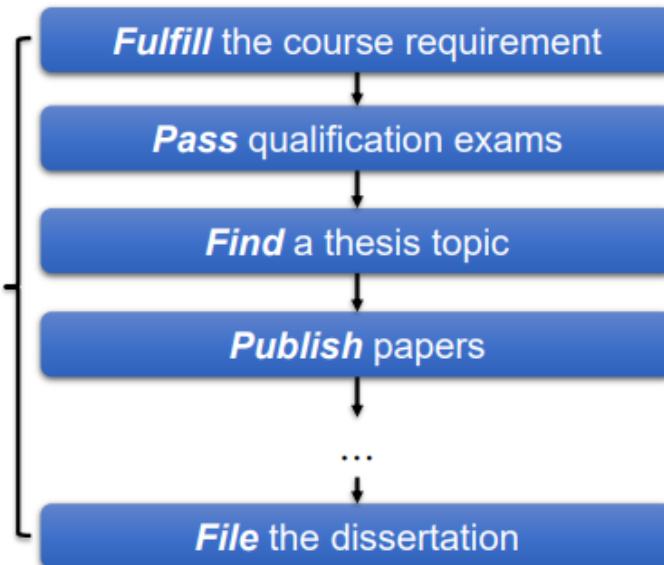
are described in different granularities,



Event Process

An event process (or event chain)

- Partially ordered events that are centered around common protagonists [Chambers et al., ACL-08]



Prediction problems on event processes

Event process completion

- What happens next?

Intention prediction

- What is the goal of “digging a hole, putting some seeds in the hole and filling it with soil”?

Membership prediction

- What are the steps of “buying a car”?

Salience prediction

- Is *defending the dissertation* more important than *doing an internship*?

Predicting Events

1. Predicting steps of the process



2. Inducing the entire process from scratch.



Predicting Events

Chambers and Jurafsky. Unsupervised Learning of Narrative Event Chains. ACL-08

Unsupervised event process completion can be done using corpus statistics (Gigaword in this work)

- Capturing the co-occurrence of events using pointwise mutual information

$$pmi(e(w, d), e(v, g))$$

- The next most likely forthcoming event can be found by maximizing the accumulated PMI

$$\max_{j:0 < j < m} \sum_{i=0}^n pmi(e_i, f_j)$$

(n : #events in the process; m : #events in the vocabulary).

Known events:

(pleaded subj), (admits subj), (convicted obj)

Likely Events:

sentenced obj	0.89	indicted obj	0.74
paroled obj	0.76	fined obj	0.73
fired obj	0.75	denied subj	0.73

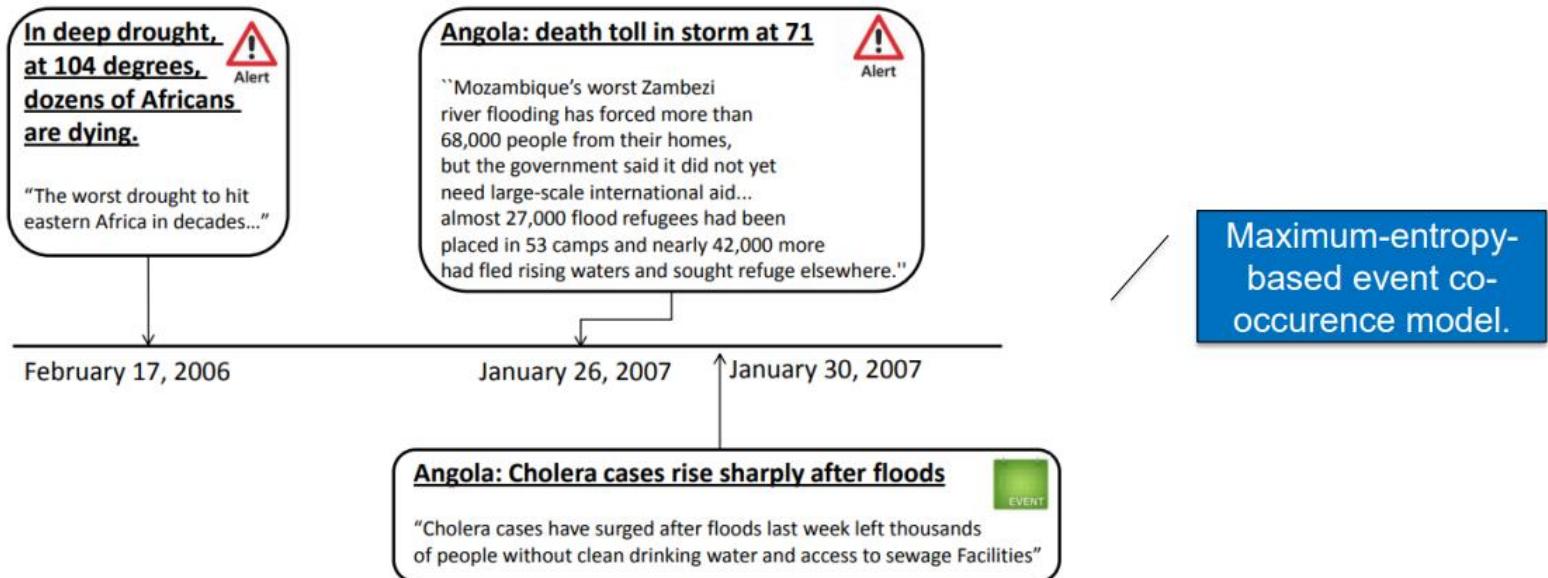


Improves narrative cloze tests (36% improvement on NYT Narrative Cloze).

Predicting Events

Radinsky and Horvitz. Mining the Web to Predict Future Events. WSDM, 2013

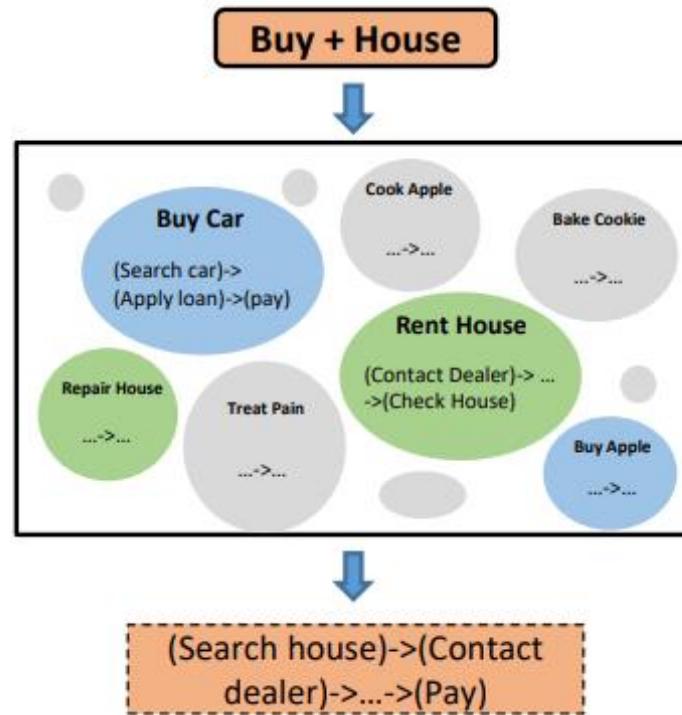
Extension of the event chain model on multiple **dated** and **topically cohesive** documents.



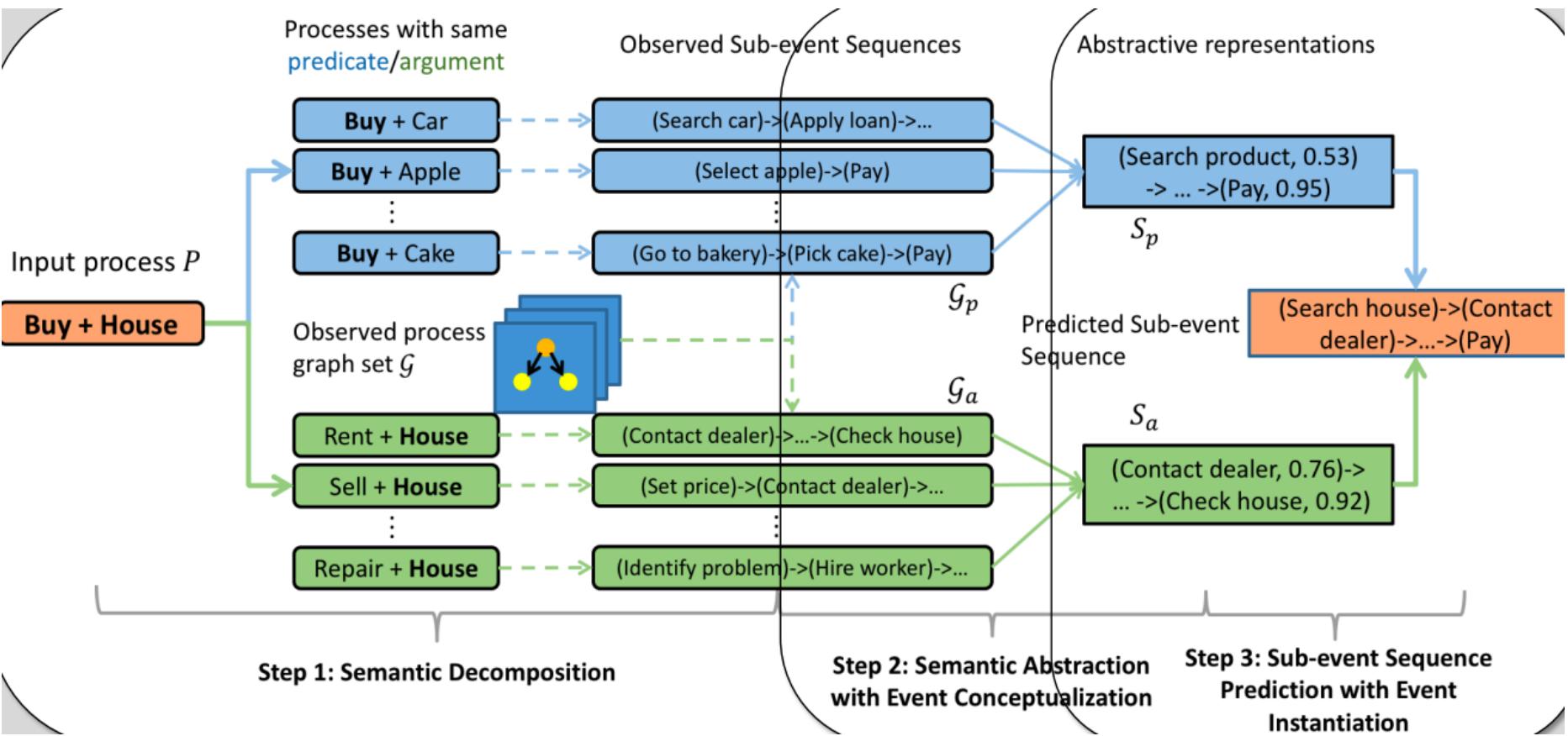
The likelihood of **cholera rising** is predicted **high** after a drought followed by storms in Angola (*based on corpus statistics*).

Predicting Events

Zhang, et al. Analogous Process Structure Induction for Sub-event Sequence Prediction. EMNLP, 2020

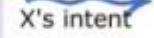


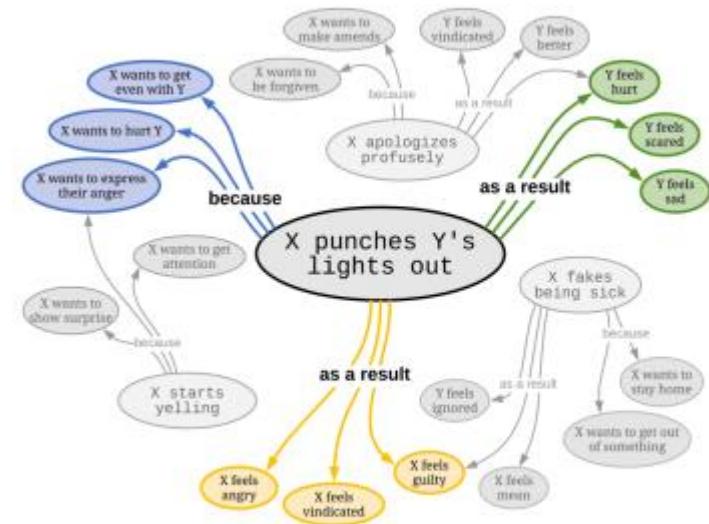
Predicting Events



Predicting Event Intention

People can easily anticipate the intents and possible reactions of participants in an event.

PersonX cooks thanksgiving dinner	 X's intent  X's reaction  Y's reaction	to impress their family tired, a sense of belonging impressed
PersonX drags PersonX's feet	 X's intent  X's reaction  Y's reaction	to avoid doing things lazy, bored frustrated, impatient



A commonsense-aware system should also perform such prediction.

Event2Mind – A learning system that understands stereotypical intents and reactions to events (Rashkin et al. ACL-18)

Predicting Event Intention

Is developed based on large crowdsourced corpora:

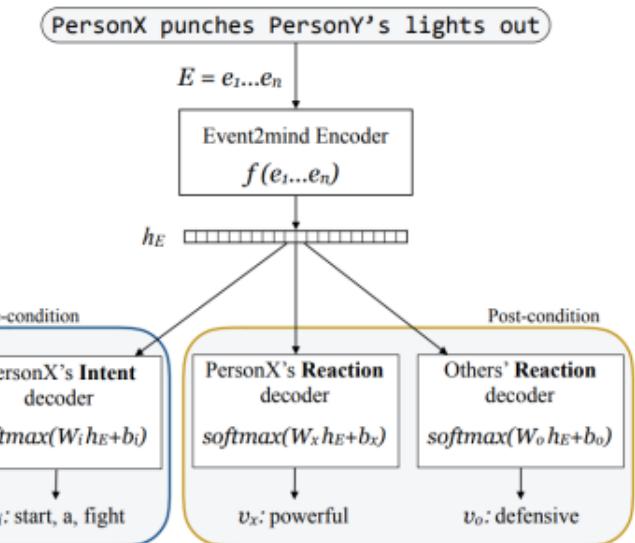
- 25,000 events
- Free-form descriptions of their intents and reactions

Performs Seq2NGram generation:

PersonX's intent: ["steak", "to kill their hunger", "to make dinner for the family", "to eat steak"]

PersonX's reaction: ["excited", "accomplished", "proud", "full"]

Other people's reaction: ["none", "happy", "person x cooked well."]

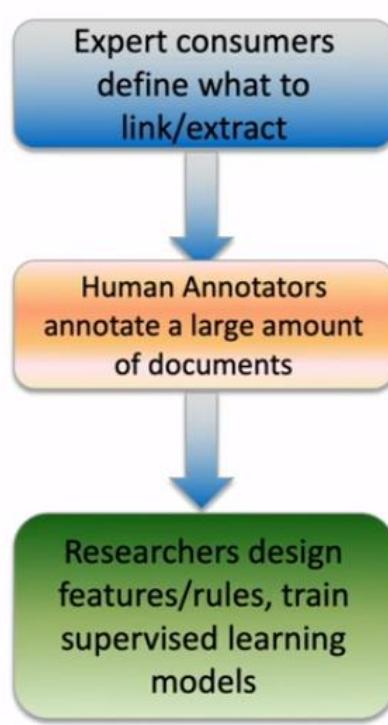


More follow-ups of Event2Mind

- ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning (Sap+ AAAI 2019)
- COMET: Commonsense Transformers for Automatic Knowledge Graph Construction (Bosselut+, ACL-19)

Cross lingual Event Extraction

English IE: Expensive but Generally Happy



- *High Cost:* requires manual clean annotations for 500 documents
- *Poor Portability:* e.g., only covers 41 relation types and 33 event types
- Limited to a certain domain, genre, language, and data modality

Cross lingual Event Extraction

- 3000+ living languages, 300+ languages have digital news data
- Certain information is often reported predominantly in local news in low-resource languages
 - e.g., the vast majority of Physical-Located relations and Meeting events involving Aung San Suu Kyi are only reported locally in Burmese news
 - e.g., language barrier was one of the main difficulties faced by humanitarian workers responding to the Ebola crisis in 2014
- Publicly available gold-standard annotations for IE exist for only a few languages
- Annotations for edge (relation and event) extraction are more expensive than node (entity) extraction because relations/events are structured and require a rich label space – not suitable for crowd-sourcing



Cross lingual Event Extraction



Enhance Quality with deep knowledge
acquisition and reasoning

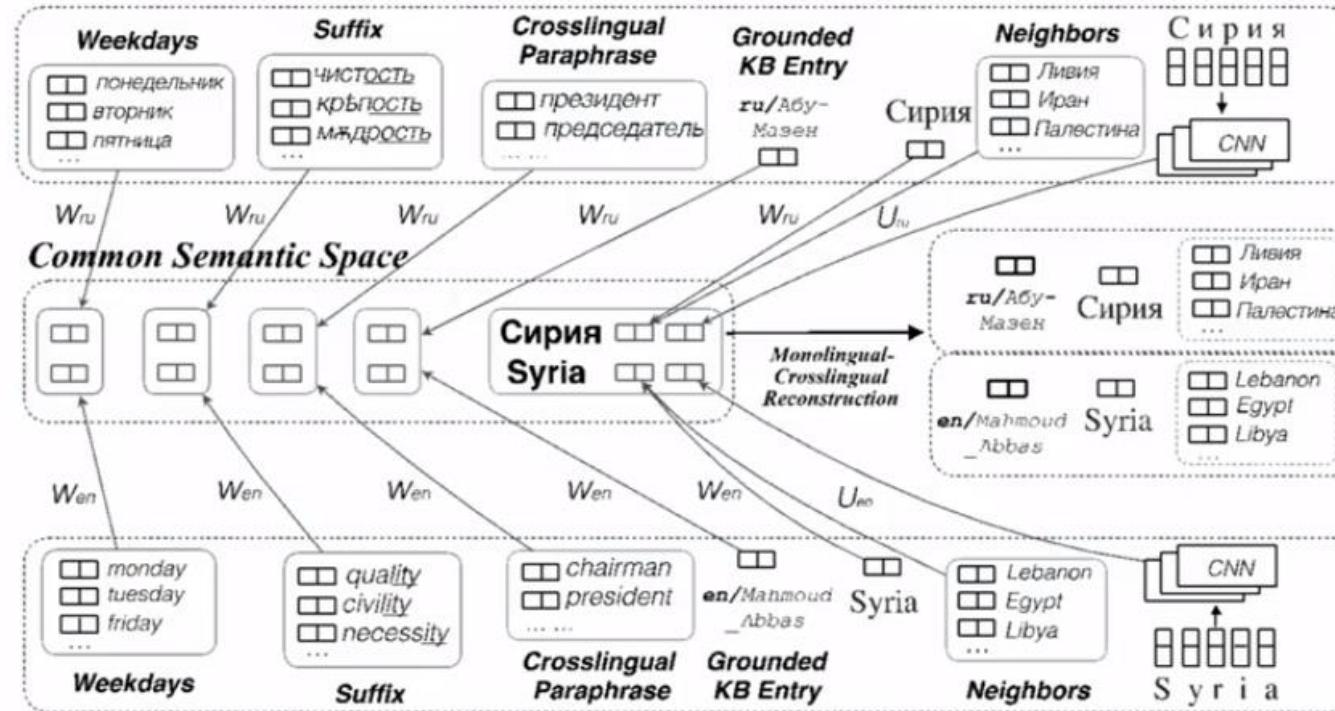


Transfer knowledge across
domain/genre/language/d
ata modality



Cross lingual Event Extraction

Russian



English

- (Huang et al., EMNLP2018)
- Our new hypothesis: Cluster distribution tends to be consistent across languages (Huang et al., EMNLP2018) ¹

Cross lingual Event Extraction

- Leverage Language-Universal Non-Conventional Linguistic Resources
 - Cross-lingual Embedding Representations
 - Cluster-consistent embedding: avoid using bi-lingual dictionaries or parallel corpora
 - Joint Entity and Word embedding
 - Cross-lingual language modeling for contextualized embedding
 - Cross-lingual Transfer Learning
 - Multi-task Multi-lingual transfer learning
 - Adversarial learning to select language-universal resources and features
 - Allow non-speakers to annotate any language
-

References

- Slide sources
 - Most of the slides today came from a lecture of Dan Jurafsky's in Chris Manning and Dan Jurafsky's online NLP course at Stanford (covers very broad range of NLP and Machine Learning topics)
 - <https://www.youtube.com/@farshadnoravesh/videos>
 - <https://www.youtube.com/watch?v=MLITKOKIHY0>
 - <https://www.youtube.com/watch?v=vZtWTzoDYXU&t=54s>
 - <https://towardsdatascience.com/from-text-to-knowledge-the-information-extraction-pipeline-b65e7e30273e>

References

<https://spacy.io/usage/v3>

<https://github.com/davidsbatista/Annotated-Semantic-Relationships-Datasets>

<https://www.youtube.com/watch?v=qpa99XSHak>

<https://slideplayer.com/slide/7234973/>

https://nlpprogress.com/english/relationship_extraction.html

<https://slideplayer.com/slide/5802639/>

<https://www.youtube.com/watch?v=18CTdWcJGL0>

<https://web.stanford.edu/~jurafsky/slp3/>

<https://slideplayer.com/slide/3367997/>

<https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html>

<https://aclanthology.org/2021.acl-long.344.mp4>



Natural Language Processing Applications



BITS Pilani
Pilani Campus

Dr. Chetana Gavankar, Ph.D,
IIT Bombay-Monash University Australia
Chetana.gavankar@pilani.bits-pilani.ac.in



Session 14: Sentiment Analysis

Date – 24th March 2024

Time – 1.40 pm to 3.40 pm

These slides are prepared by the instructor, with grateful acknowledgement of Prof. Dan Jurafsky and many others who made their course materials freely available online.

Outline

- Motivation
 - What is sentiment analysis
 - Why sentiment analysis
 - Sentiment analysis methods
 - Sentiment lexicons
 - Methods for learning sentiment lexicons
 - Aspect based sentiment analysis
 - Opinion spamming
-

Motivation For Sentiment Analysis



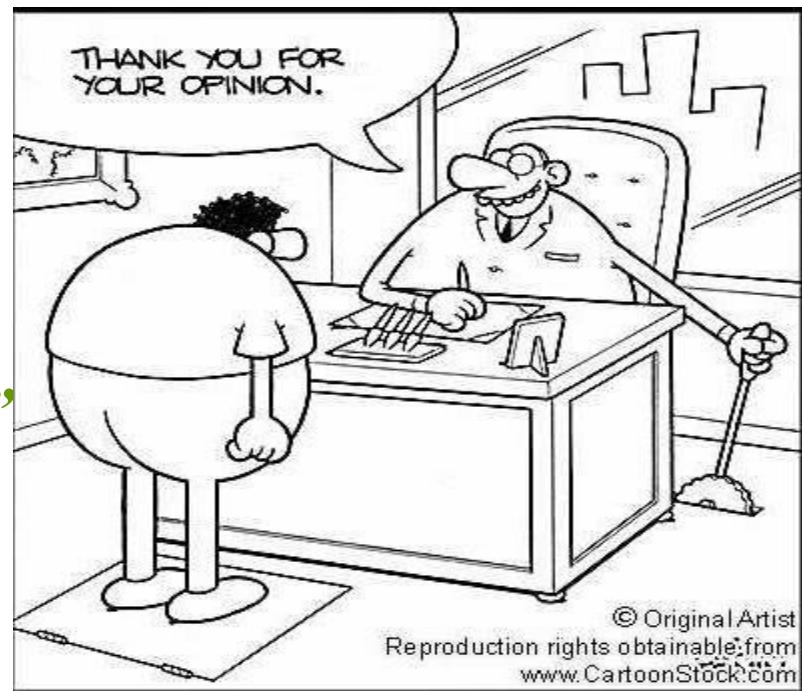
What others think has always been an important piece of information

“Which car should I buy?”

“Which schools should I apply to?”

“Which Professor to work for?”

“Whom should I vote for?”



“So whom shall I ask?”

Pre Web

- Friends and relatives
- Acquaintances
- Consumer Reports



Post Web

“...I don’t know who..but apparently it’s a good phone. It has good battery life and...”

- Blogs (google blogs, livejournal)
- E-commerce sites (amazon, ebay)
- Review sites (CNET, PC Magazine)
- Discussion forums (forums.craigslist.org,
forums.macrumors.com)
- Friends and Relatives (occasionally)



The problem is..

- “Whoala! I have the reviews I need”
 - *Now that I have “**too much**” information on one topic...I could easily form my opinion and make decisions...*
 - Is this true?
 - ...Not Quite
 - Searching for reviews may be difficult
 - Can you search for opinions as conveniently as general Web search?
 - eg: is it easy to search for “*iPhone vs Google Phone*”?
-

Facts and Opinions

Two main types of information on the Web.

- Facts(Objective) and Opinions(Subjective)

Fact : Thursday is a day.

Opinion : Thursday was a fun day.

Fact : iPhone is an Apple product.

Opinion : iPhone is good.

- Google searches for facts (currently)
 - Facts can be expressed with topic keywords
 - Google does not search for opinions
 - Opinions are hard to express with keywords
-

Issues

- Not all subjective sentences contain opinions, e.g.
 - *"I want a phone with good voice quality"*
- Not all objective sentences contain no opinions, e.g.
 - *"The earphone broke in just two days!"*

What is Sentiment analysis

- Computational study of opinions, sentiments, evaluations, attitudes, appraisal, affects, views, emotions, subjectivity, etc., expressed in text.



Sentiment Analysis



It's a big day & I want to book a table at a nice Japanese restaurant

Seattle has many
★★★★★
sushi restaurants



What are people
saying about
the food?
the ambiance?...



Sentiment Analysis



Positive reviews not positive about everything

Sample review:

Watching the chefs create incredible edible art made the experience very unique.

My wife tried their ramen and it was pretty forgettable.

All the sushi was delicious! Easily best sushi in Seattle.

Experience



Sentiment Analysis



From reviews to topic sentiments

All reviews
for restaurant

★★★★★ 10/10/2015
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese eel (unagi), and the signature tofu and 2 special rolls. I would skip the special rolls, because the scallop was delicious, but the eel was a bit too salty. The eel, as recommended by other Yelpers was amazing. It's more chewy and the soy-sesame-ginger is the perfect amount of flavor for the delicate tofu.

★★★★★ 01/10/2015
Dining here at the sushi bar inside the flat like setting front row to an amazing performance. The food I have eaten, ordered during the 10 after work, got here breathlessly of 8:15pm, and got the last two seats in the place.

★★★★★ 04/05/2015
I came here having high expectations due to the reviews of this place, but I was left disappointed. This restaurant is small so do make reservations when you come here. Sushi cost from \$4-10 each and dishes are small.

Novel intelligent
restaurant review app

Experience
★★★★★

Ramen
★★★

Sushi
★★★★★

Easily best sushi
in Seattle.

Examples



The screenshot shows the top navigation bar of the Amazon.com website. It includes the Amazon logo with 'prime' underneath, a search bar with a magnifying glass icon, and various top-level links: Buy Again, Browsing History, Cody's Amazon.com, Early Black Friday Deals, Gift Cards, Registry, Sell, and Help.

LG Electronics OLED55E8PUA 55-Inch 4K Ultra HD Smart OLED TV (2018... > Customer reviews

Customer reviews

★★★★★ 18

3.9 out of 5 stars



LG Electronics OLED55E8PUA 55-Inch 4K Ultra HD Smart OLED TV (2018 Model)

by LG

Size: 55-inch Change

Price: \$2,296.99  prime

[Write a review](#)

Top positive review

[See all 14 positive reviews](#) >



Mayra S. TOP 1000 REVIEWER

★★★★★ With Google Assistant and new Alpha 9 Processor, 2018 LG Oled's are great upgrades for first time 4K/HDR/Oled Owners

May 3, 2018

Top critical review

[See all 4 critical reviews](#) >



Brett W.

★★★★☆ Extreme stuttering (no soft transition between frames) is an important factor to consider with OLED TV's

August 3, 2018

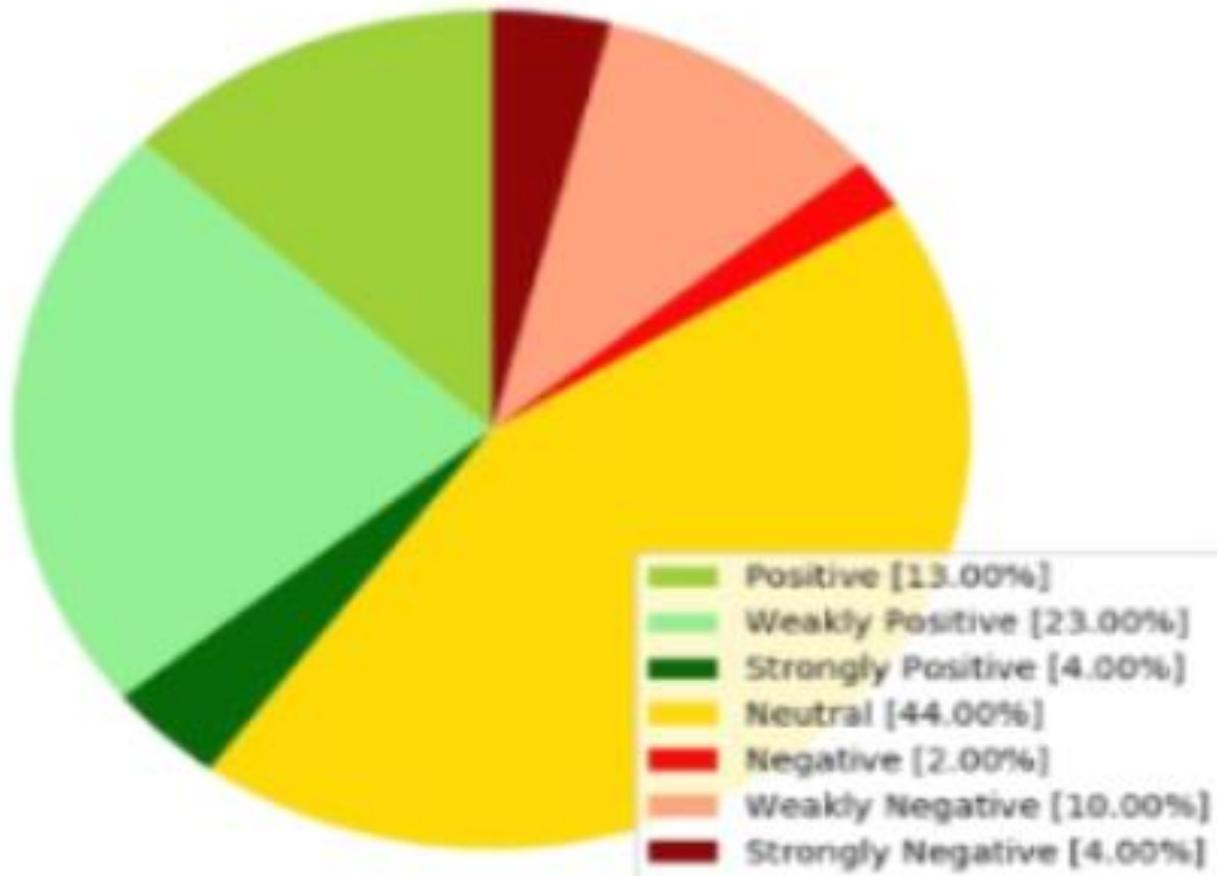
Examples

Data gathered from the analysis of +4,000 product reviews



Examples

How people are reacting on bitcoin by analyzing 100 Tweets.



Examples

Sentiment Analytics for the Telecom Company

Negative review

Dear #XYZ there is no network in my area and internet service is pathetic from the past one week. Kindly help me out

-Dated: 10/09/17

Mixed review

Although the value added services being provided are great but the prices are high
#VAS #XYZ

-Dated: 10/09/17

Positive review

Great work done #XYZ Problem resolved by customer care in just one day
#ThanksXYZ

-Dated: 5/06/17



Sentiment Analytics Model

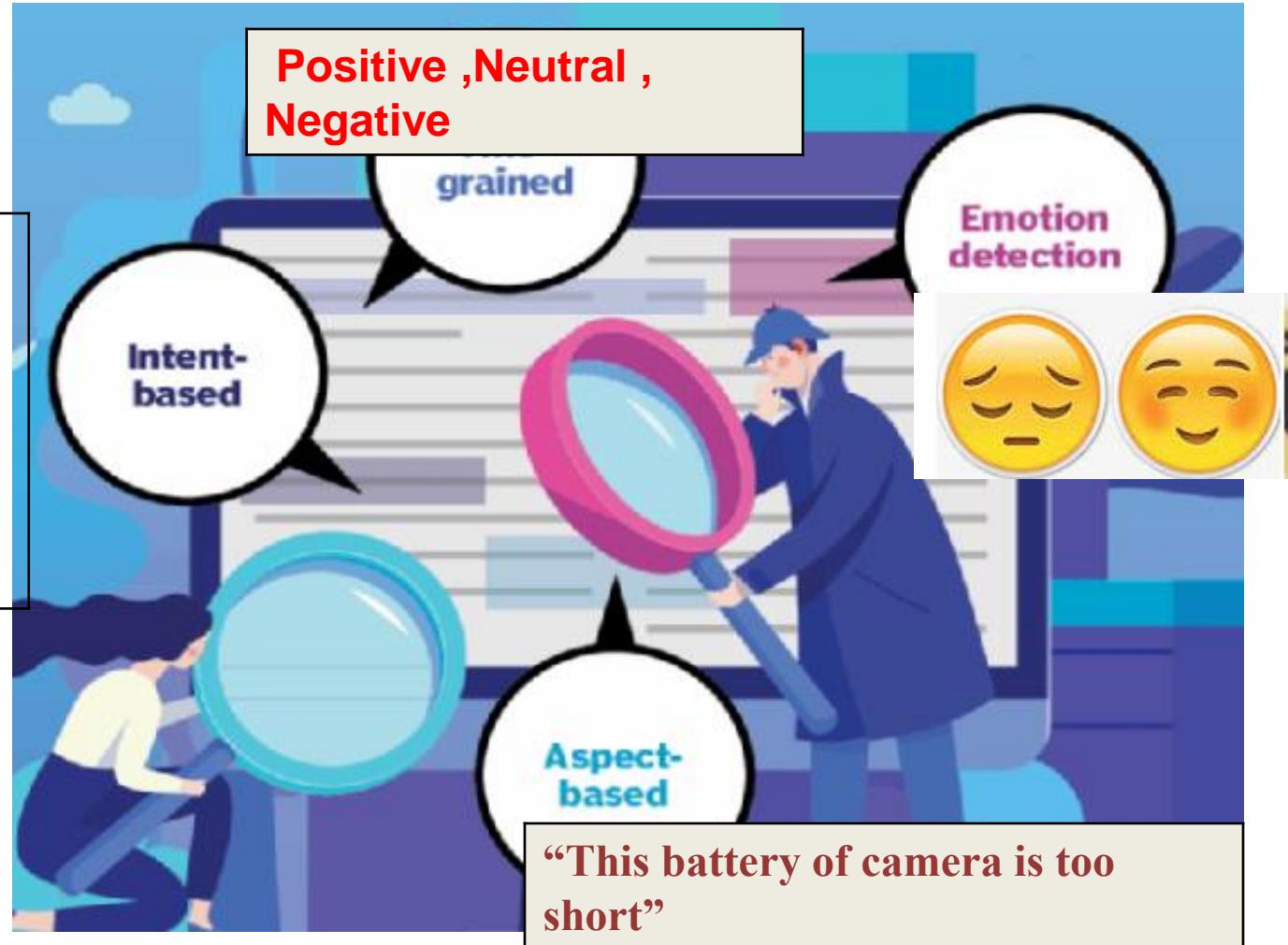
Analyze output to implement **Actionables** like:

- Improve service quality and increase quality checks for locations with max negative reviews.
- Train employees and improve infrastructure of areas with poor service.
- Ensure dedicated customer complaint teams for some areas.
- Tweak the marketing strategy to convey the right message to the customers

* XYZ depicts name of the telecom company

Types of Sentiment Analysis

intention to sell,
intention to complain or intention to purchase etc

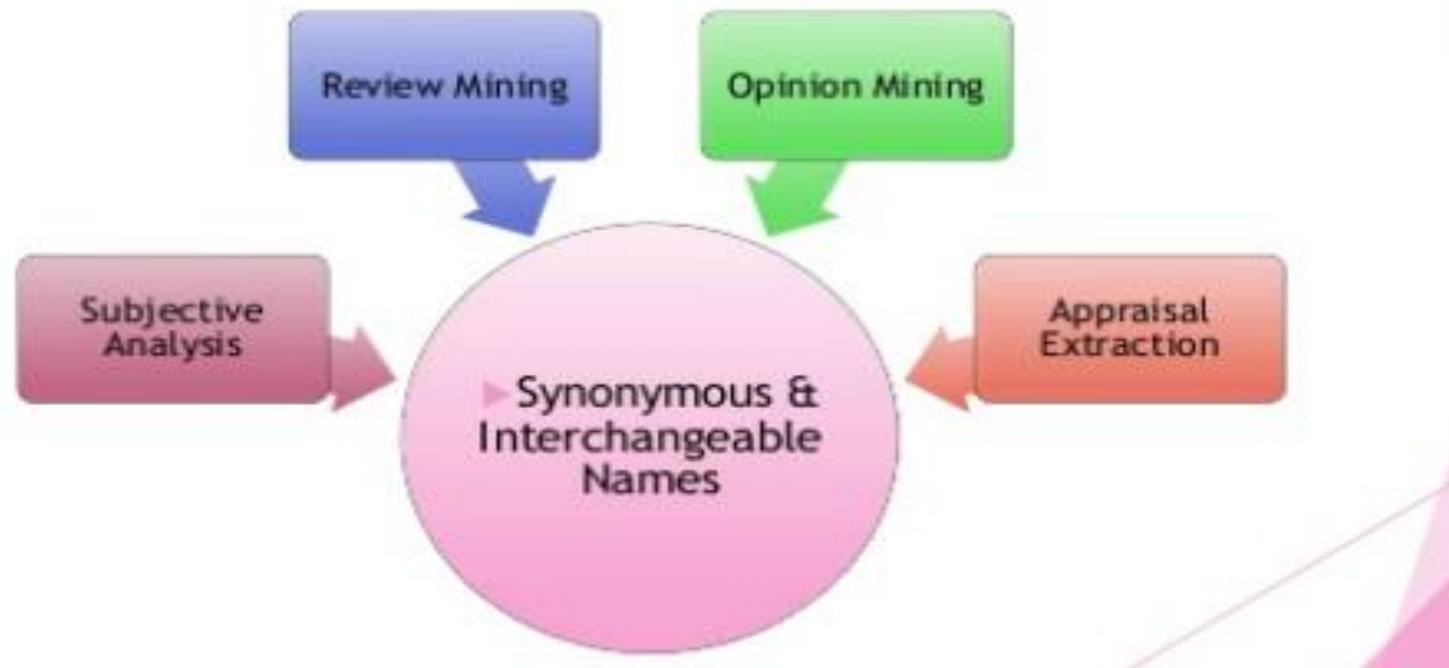




Different levels of sentiment analysis

- Three levels of granularity
 - Document level
 - Sentence level
 - Entity and Feature/Aspect level

Different terms for sentiment analysis



Sentiment analysis methods

- **Rule-based** systems that perform sentiment analysis based on a set of manually crafted rules.
 - **Automatic** systems that rely on machine learning techniques to learn from data.
 - **Hybrid** systems that combine both rule-based and automatic approaches.
-

Rule based methods

Following steps need to be performed

- Extract the data
 - Tokenize text. The task of splitting the text into individual words
 - Stop words removal. Those words which do not carry any significant meaning and should not be used for the analysis activity. Examples of stop words are: a, an, the, they, while etc.
 - Punctuation removal (in some cases)
 - Running the *preprocessed* text against the sentiment lexicon which should provide the number/measurement corresponding to the inferred emotion
-

Example

“Sam is a great guy”

1. Tokenize

“*Sam is a great guy.*”



Tokenize
Sam
is
a
great
guy
.

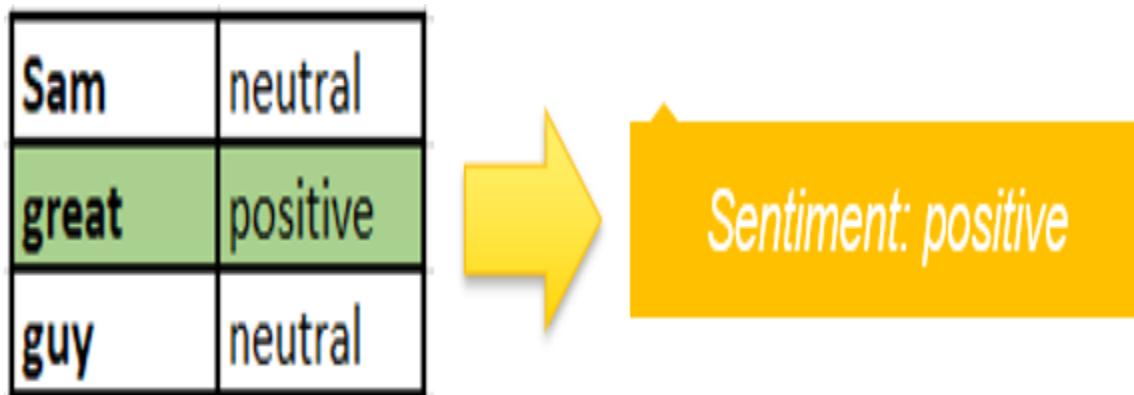
2. Remove stop words and punctuations

Tokenize	Preprocessing
Sam	
is	stop word
a	stop word
great	
guy	
.	punctuation



Sam	neutral
great	positive
guy	neutral

-
3. Running the lexicon on the preprocessed data, returns a **positive sentiment** score/measurement because of the presence of a positive word “great” in the input data.



Machine learning Approach

The song was good .

1.Tokenization

- The
- Song
- Was
- Good
- .

2.Cleaning the data (Remove special characters)

- The
- Song
- Was
- Good

Contd..

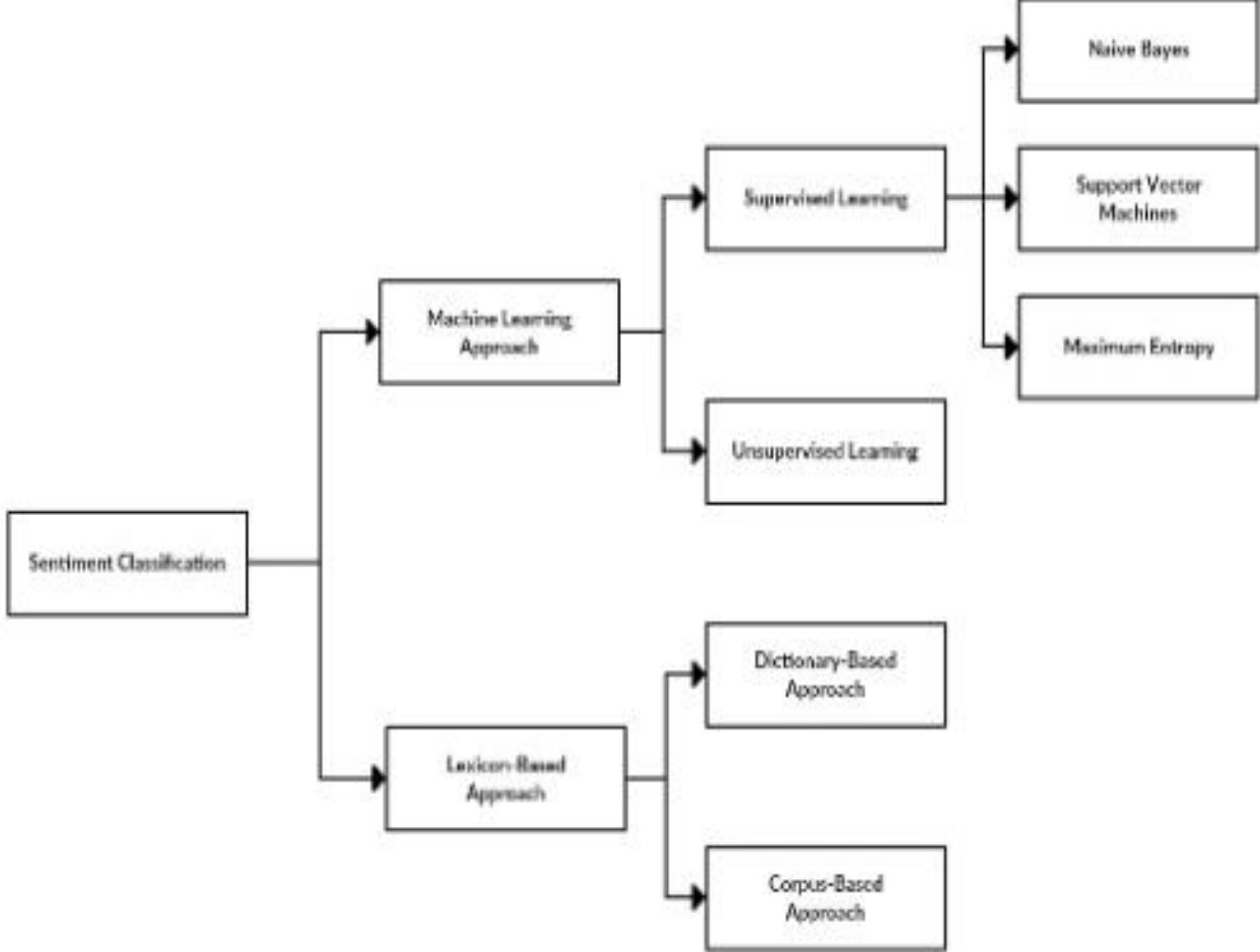
3. Remove stop words.

- Song
- good

4. Classification(Positive, negative ,Neutral)

Apply supervised algorithm

- *Naïve Bayes*
 - *Support vector machines*
 - *Maximum Entropy*
-



Sentiments/ Emotions

- Most models include the two dimensions valence and arousal, and many add a third, dominance.
 - These can be defined as:
 - valence: the pleasantness of the stimulus
 - arousal: the intensity of emotion provoked by the stimulus
 - dominance: the degree of control exerted by the stimulus
-

Lexicons

- Many sentiment applications rely on lexicons to supply features to a model.
 - A lexicon is a **resource with information about words**.
 - A sentiment lexicon has information such as list of words which are positive and negative.
-

General Inquirer

- Harvard General Inquirer Database (Stone, 1966)
 - Total of 11,788 terms
 - http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
 - <http://www.wjh.harvard.edu/~inquirer/homecat.htm>
 - Positive (1915 words) vs Negative (2291 words)
 - Strong vs Weak
 - Active vs Passive
 - Overstated versus Understated
 - Pleasure, Pain, Virtue, Vice
 - Motivation, Cognitive Orientation, etc

Sample

A fragment of the Harvard General Inquirer spreadsheet file.

	Entry	Positiv	Negativ	Hostile	...184 classes ...	Othtags	Defined
1	A					DET ART	...
2	ABANDON		Negativ			SUPV	
3	ABANDONMENT		Negativ			Noun	
4	ABATE		Negativ			SUPV	
5	ABATEMENT					Noun	
...							
35	ABSENT#1		Negativ			Modif	
36	ABSENT#2					SUPV	
...							
11788	ZONE					Noun	

SentiWordNet

- Home page :<http://sentiwordnet.isti.cnr.it/>
 - All WordNet synsets automatically annotated for degrees of positivity ,negativity and neutrality/objectiveness.
-

Example

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	able#1	(usually followed by 'to') having the necessary means or [...]
a	00002098	0	0.75	unable#1	(usually followed by 'to') not having the necessary means or [...]
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ or organism; [...]
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis of an organ or organism; [...]
a	00002730	0	0	acrosopic#1	facing or on the side toward the apex
a	00002843	0	0	basiscopic#1	facing or on the side toward the base
a	00002956	0	0	abducting#1 abducent#1	especially of muscles; [...]
a	00003131	0	0	adductive#1 adducting#1 adducent#1	especially of muscles; [...]
a	00003356	0	0	nascent#1	being born or beginning; [...]
a	00003553	0	0	emerging#2 emergent#2	coming into existence; [...]



MPQA Subjectivity Cues Lexicon

- <https://mpqa.cs.pitt.edu/>
- 6885 words from 8221 lemmas
 - 2718 positive
 - 4912 negative
- Each word annotated for intensity (strong, weak)

	Strength	Length	Word	Part-of-speech	Stemmed	Polarity
1.	type=weaksubj	len=1	word1=abandoned	pos1=adj	stemmed1=n	priorpolarity=negative
2.	type=weaksubj	len=1	word1=abandonment	pos1=noun	stemmed1=n	priorpolarity=negative
3.	type=weaksubj	len=1	word1=abandon	pos1=verb	stemmed1=y	priorpolarity=negative
4.	type=strongsubj	len=1	word1=abase	pos1=verb	stemmed1=y	priorpolarity=negative
5.	type=strongsubj	len=1	word1=abasement	pos1=anypos	stemmed1=y	priorpolarity=negative
6.	type=strongsubj	len=1	word1=abash	pos1=verb	stemmed1=y	priorpolarity=negative
7.	type=weaksubj	len=1	word1=abate	pos1=verb	stemmed1=y	priorpolarity=negative
8.	type=weaksubj	len=1	word1=abdicate	pos1=verb	stemmed1=y	priorpolarity=negative
9.	type=strongsubj	len=1	word1=aberration	pos1=adj	stemmed1=n	priorpolarity=negative
10.	type=strongsubj	len=1	word1=aberration	pos1=noun	stemmed1=n	priorpolarity=negative
...						
8221.	type=strongsubj	len=1	word1=zest	pos1=noun	stemmed1=n	priorpolarity=positive

Linguistic inquiry and word count

- Home Page: <http://www.liwc.net/>
 - 2300 word > 70 classes
 - Affective Processes
 - Negative emotion (bad, weird, hate, problem,tough)
 - Positive emotion (love,nice,sweet)
 - Cognitive Processes
-

Category	Examples
Negate	aint, ain't, arent, aren't, cannot, cant, can't, couldnt, ...
Swear	arse, arsehole*, arses, ass, asses, asshole*, bastard*, ...
Social	acquainta*, admit, admits, admitted, admitting, adult, adults, advice, advis*
Affect	abandon*, abuse*, abusi*, accept, accepta*, accepted, accepting, accepts, ache*
Posemo	accept, accepta*, accepted, accepting, accepts, active*, admir*, ador*, advantag*
Negemo	abandon*, abuse*, abusi*, ache*, aching, advers*, afraid, aggravat*, aggress*,
Anx	afraid, alarm*, anguish*, anxi*, apprehens*, ashame*, aversi*, avoid*, awkward*
Anger	jealous*, jerk, jerked, jerks, kill*, liar*, lied, lies, lous*, ludicrous*, lying, mad

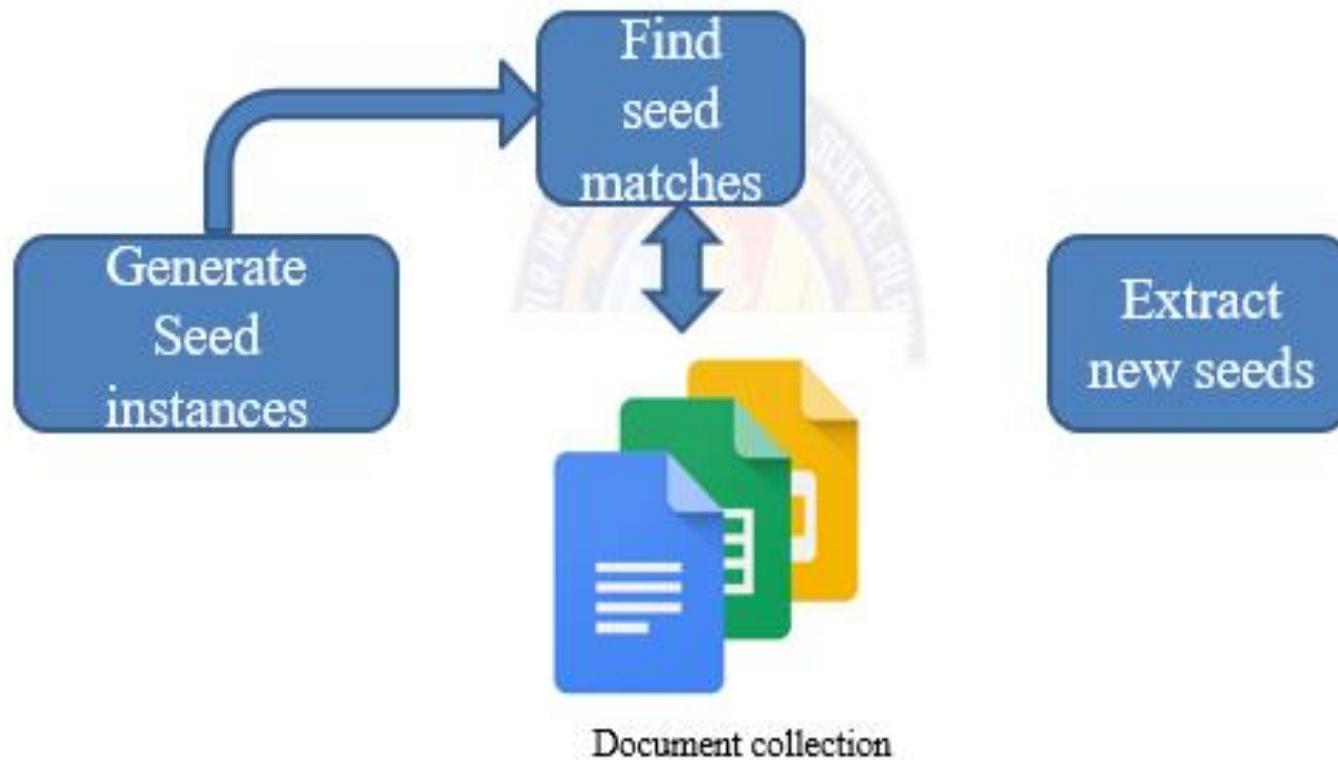
Bing Liu Opinion Lexicon

- [Bing Liu's Page on Opinion Mining](#)
 - <http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>
 - 6786 words
 - 2006 positive
 - 4783 negative
-

Corpus based lexicon generator

- A more sophisticated technique is a corpus-based approach which relies on syntactic or **co-occurrence patterns** together with a seed list of opinion words.
 - The technique **starts with a list of seed opinion adjective words**, and uses them and a **set of linguistic constraints** or conventions on connectives to **identify additional adjective opinion words and their orientations**.
-

Bootstrapping architecture



Example

- Adjectives conjoined by “and” have same polarity

Fair and legitimate ,corrupt and brutal

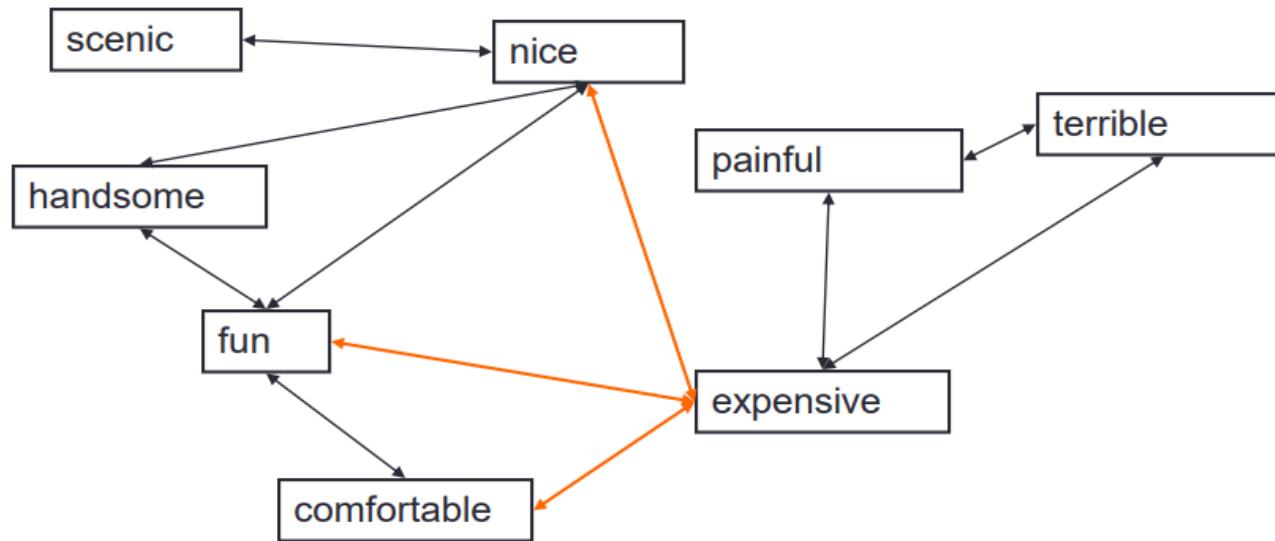
- Adjectives conjoined by “but” do not

Fair but brutal



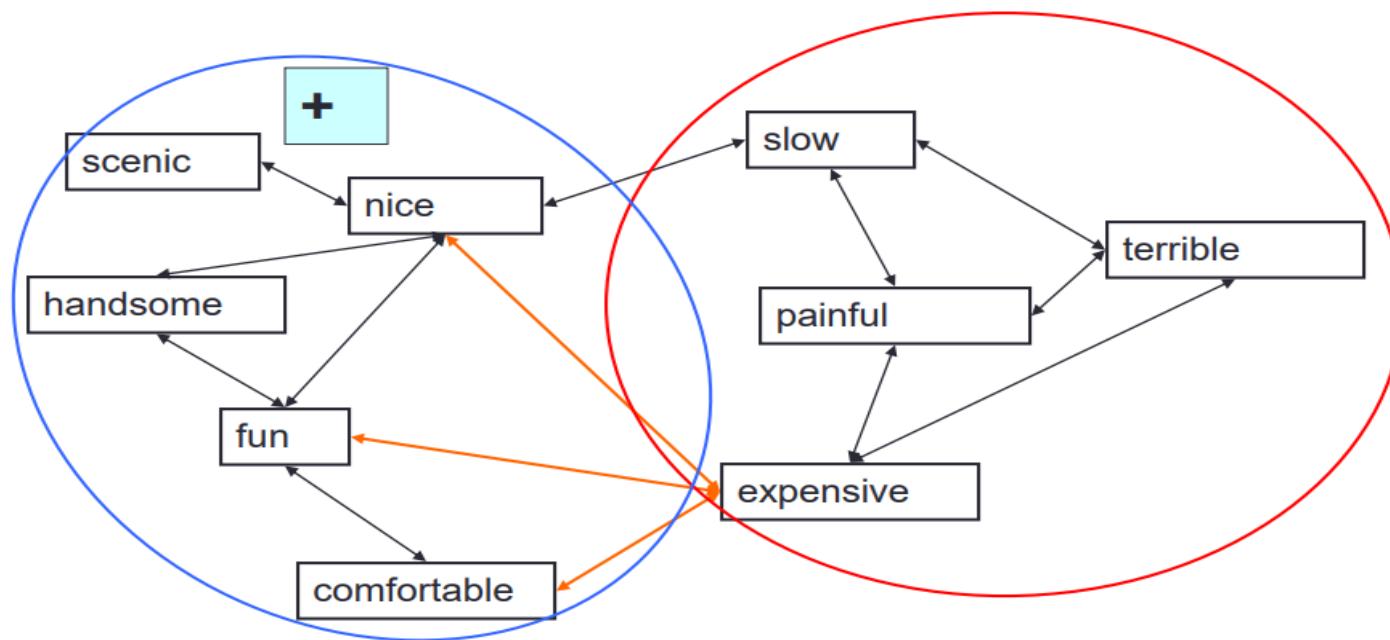
Algorithm

1. Generate a Labeled seed set of adjectives
2. Expand seed set to conjoined adjectives by looking up in a corpus/web search
3. builds a graph of adjectives linked by the same or different semantic orientation



Clustering Algorithm

- A clustering algorithm partitions the adjectives into two subsets



Turney algorithm

- Extract a phrasal lexicon from reviews
- Learn polarity of each phrase
- Rate a review by the average polarity of its phrases

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

Two-word phrases with adjectives

How to measure polarity of a phrase

- Positive phrases co-occur more with “excellent”
 - Negative phrases co-occur more with “poor”
 - But how to measure co-occurrence?
-

Pointwise Mutual Information

- Pointwise mutual information: How much more do events x and y co-occur than if they were independent?

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

- If two words are **statistically independent**, $\text{PMI}=0$
 - If two words tend to not at all **co-occur**, PMI is **negative**
 - If two words tend to co-occur, PMI is **positive**
 - Does phrase appear more with “poor” or “excellent”?
– $\text{Polarity}(\text{phrase}) = \text{PMI}(\text{phrase}, \text{"excellent"}) - \text{PMI}(\text{phrase}, \text{"poor"})$
-

Two reviews for Positive and Negative phrases



Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
...		
low fees	JJ NNS	0.33
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5
<i>Average</i>		0.32

Phrase	POS tags	Polarity
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
...		
virtual monopoly	JJ NN	-2.0
lesser evil	RBR JJ	-2.3
other problems	JJ NNS	-2.8
low funds	JJ NNS	-6.8
unethical practices	JJ NNS	-8.5
<i>Average</i>		-1.2

Wordnet based polarity estimation

- WordNet: online thesaurus indexing words by synonyms
 - Create positive (“good”) and negative seed-words (“terrible”)
 - Find Synonyms and Antonyms
 - Positive Set: Add synonyms of positive words (“well”) and antonyms of negative words
 - Negative Set: Add synonyms of negative words (“awful”) and antonyms of positive words (“evil”)
 - Repeat, following chains of synonyms
 - Filter
-

Aspect Based Sentiment Analysis (ABSA)

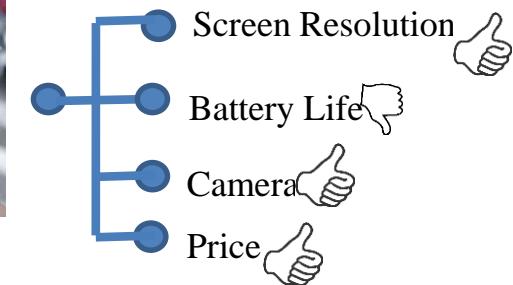


“(1) I bought an *iPhone* a few days ago. (2) It was such a *nice phone*. (3) The *touch screen* was really *cool*. (4) The *voice quality* was *clear* too. (5) Although the *battery life* was *not long*, that is ok for me. (6) However, *my mother* was mad with me as I did not tell her before I bought it. (7) She also thought the *phone* was too *expensive*, and wanted me to return it to the shop. . . ”

Aspect Based Sentiment Analysis (ABSA)

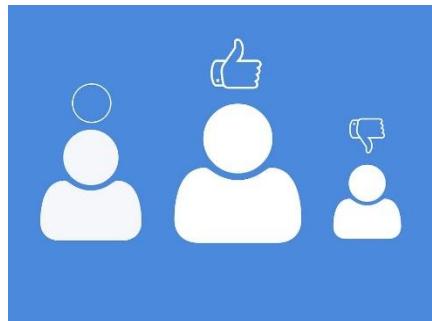


- Each opinion is defined as **quintuple** (e, a, s, h, t) , where e is an entity and a is one of its aspects, s is the sentiment on the aspect a , h is the opinion holder and t is the time when the opinion is expressed.
- Find the **target(Aspect/Entity)** of the sentiment.
- Two approaches
 - Find most common noun phrases
 - Build a classifier



Frequency-Based Aspect Extraction

- A key characteristic is that an **opinion always has a target**.
- Exploit **syntactic structures** to depict opinion and target relationships



Review corpus



Association Rule
Mining

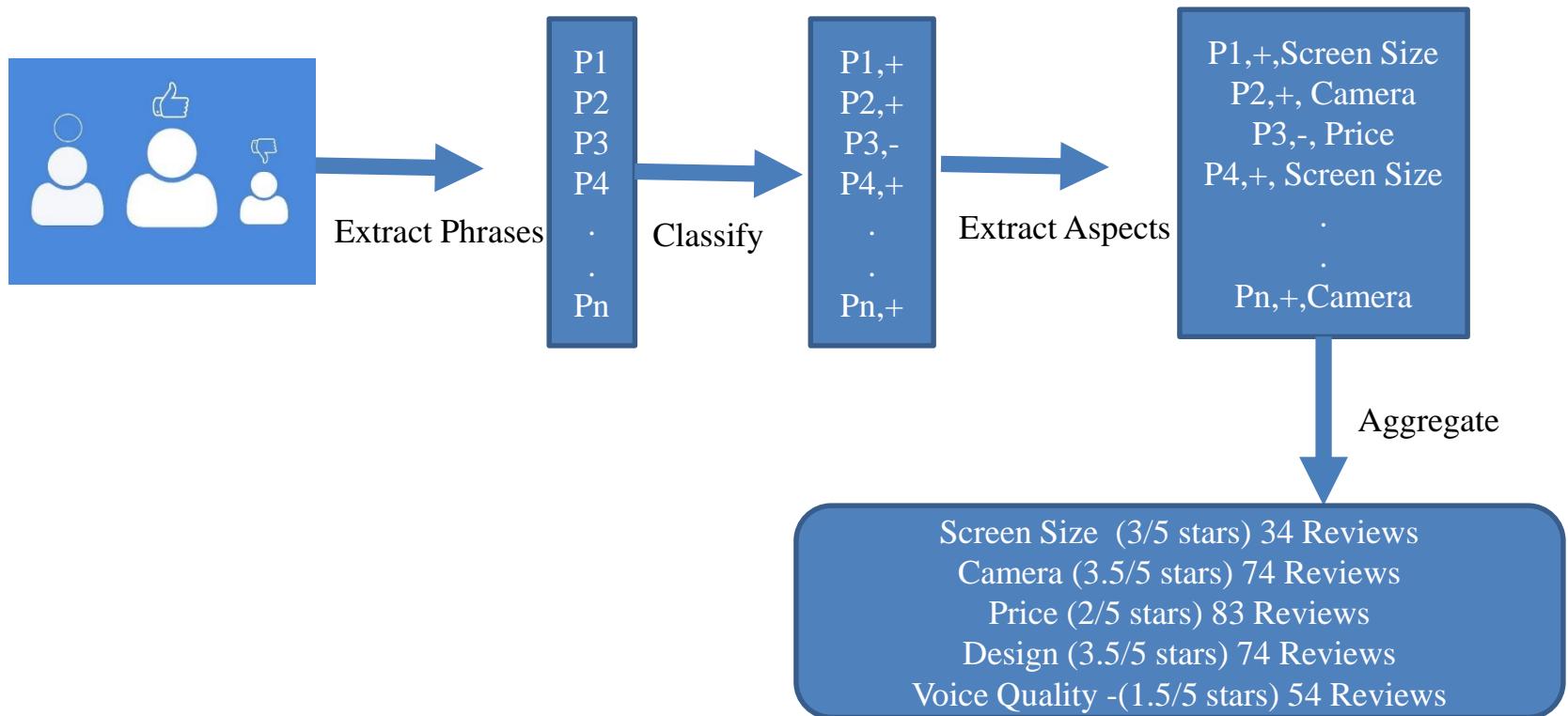
Screen Size – 100/500
Camera Resolution – 300/500
Battery Life – 350/500
Price -450/500
Voice clarity – 325/500

Examples of aspects extracted

- Those candidate aspects with the highest frequency counts are almost always the most important aspects of the product.
- Assumption: Corpus has reasonable number of reviews and belong to same product.

Entity	Aspects extracted
Casino	Casino, buffet, pool, resort, beds
Department store	Selection, department, sales, shop, clothing
Greek Restaurant	Food, Wine, Service, Appetizer, lamb

Architecture for ABSA



Blair-Glodensohn from Google

Target and Aspect Based Sentiment Analysis (TBSA)



- Finding the sentiment towards a target , its aspect and classifying the sentiment.
- Fine grained to get more information out of the text

Ex : “This book is a hardcover version, but the price is a bit high.”

Target	Aspect	Emotion
Book	Cover	Positive
Book	Price	Negative

Challenges and Approaches

- Bert is trained task agnostic - masked representation. So it is not domain specific.
- Aspect Based Sentiment Analysis is very much Domain dependent and the results depend upon the domain corpus.
- But Bert (cased) can also be further easily trained with a simple Dense layer and a soft max layer.
- Its context can be further improved as well.

Before BERT

- Target-dependent LSTM (TD-LSTM) to capture the aspect information when modeling sentences. A forward LSTM and a backward LSTM towards target words are used to capture the information before and after the aspect.
 - Attention mechanism to concentrate on corresponding parts of a sentence when different aspects are taken as input.
-



Target-Dependent Sentiment Classification With BERT

ZHENGJIE GAO , AO FENG , XINYU SONG, AND XI WU

- Traditional sentiment analysis methods require complex feature engineering and embedding representations have dominated leaderboards for a long time.
- However, the context-independent nature limits their representative power in rich context, hurting performance in Natural Language Processing (NLP) tasks.
- We implement three target-dependent variations of the BERT base model, with positioned output at the target terms and an optional sentence with the target built in.
- Dataset used - SemEval-2014 and a Twitter dataset

Conflict Emotion

- “I bought a mobile phone, its camera is wonderful but battery life is short”
 - Its camera is wonderful - positive
 - Battery life is short - negative
 - So the conveyed emotion is termed as conflict
-
- 
- 

BERT-FC and TD-BERT

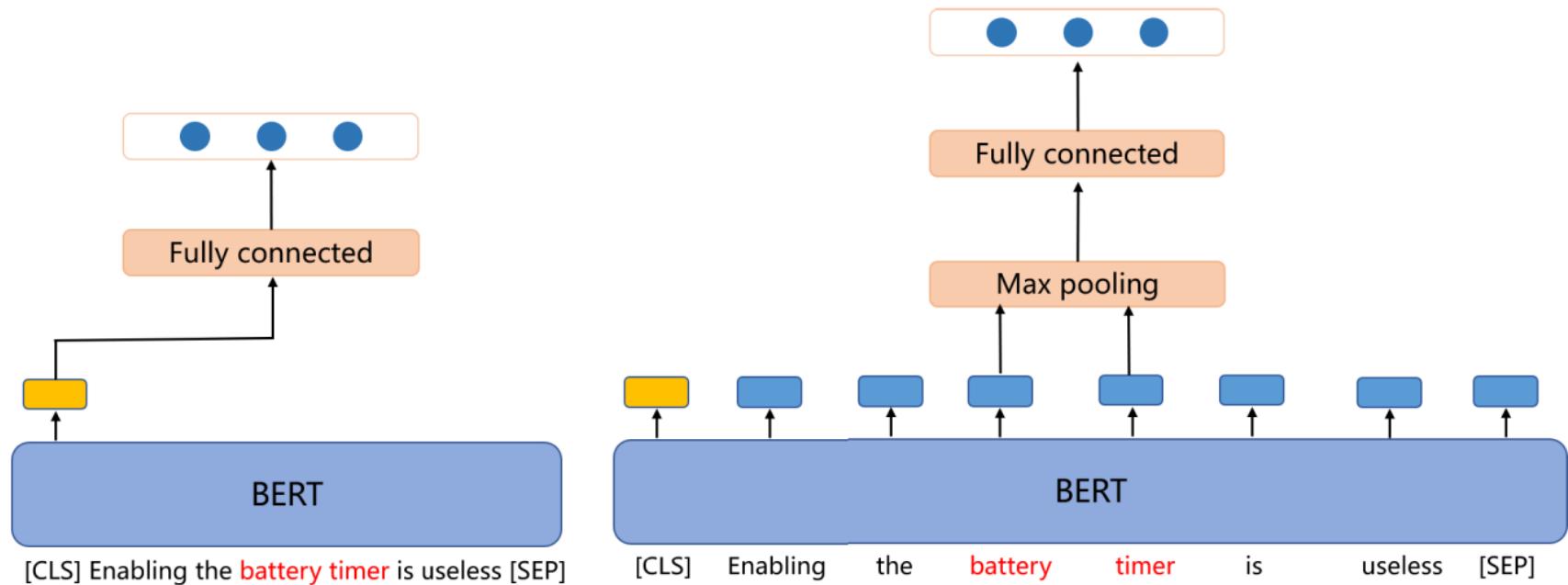


FIGURE 2. The architecture of BERT-FC (left) and TD-BERT (right).

Auxiliary Question

- Sentence - “Enabling the battery timer is useless”
- Auxiliary Question - “what do you think of the battery timer of it”
- Advantages - increases the context of the sentence, and BERT is better in Question answering when compared to sentiment analysis.

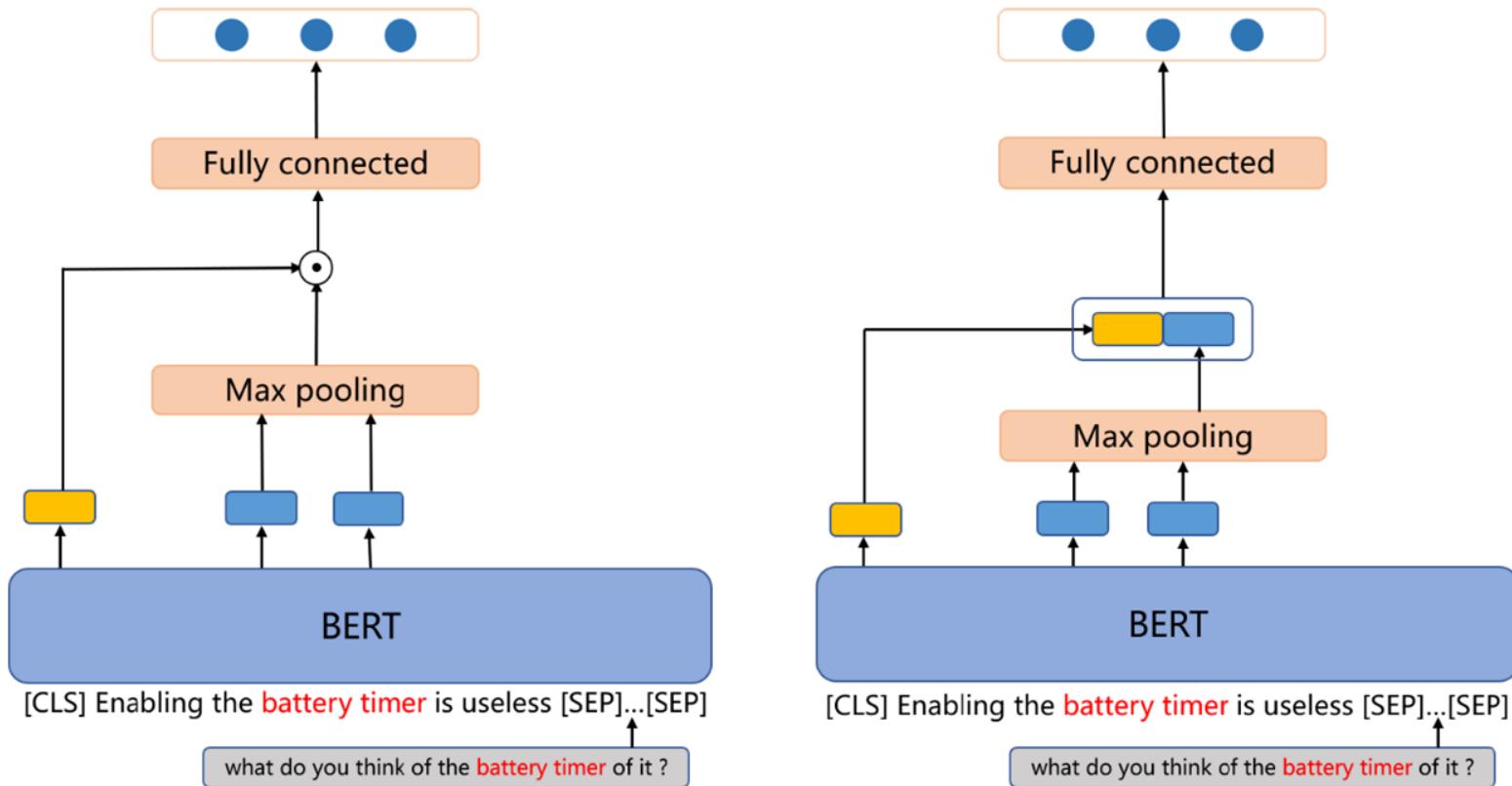


FIGURE 3. The architecture of TD-BERT-QA-MUL (left) and TD-BERT-QA-CON (right).

Dataset

TABLE 2. Statistics of the experiment datasets.

Dataset	Positive	Negative	Neutral	Conflict	Total
Laptop-Train	987	866	460	45	2358
Laptop-Test	341	128	169	16	654
Laptop-Train-Hard	159	147	173	17	496
Laptop-Test-Hard	31	25	49	3	108
Restaurant-Train	2164	805	633	91	3693
Restaurant-Test	728	196	196	14	1134
Restaurant-Train-Hard	379	323	293	43	1038
Restaurant-Test-Hard	92	62	83	8	245
Twitter-Train	1561	1560	3127	-	6248
Twitter-Test	173	173	346	-	692

Results

- Achieve new state-of-the-art performance on three datasets, especially for Twitter, in which our model has a 2-3% margin over the best previous result.
- After the position output information of the target is integrated into the BERT-pair-QA-M model, the classification accuracy of TD-BERT-QA-MUL and TD-BERT-QA-CON is also improved, slightly over TD-BERT on Twitter and Restaurant in its 3-way classification task
- The information fusion is applied with either element-wise multiplication or concatenation, but the performance comparison between them is almost equivalent.

Utilizing Bert for ASBA via constructing an auxiliary sentence

- Construct Auxiliary sentence from the aspect
- Convert ABSA to Question Answering and Natural language Inference task
- Experiment with 4 types of Auxiliary sentence
- Dataset : Sentihood and SemEval 2014

Example:

LOCATION2 is central London so extremely expensive, **LOCATION1** is often considered the coolest area of London.

Target	Aspect	Sentiment
LOC1	general	Positive
LOC1	price	None
LOC1	safety	None
LOC1	transit-location	None
LOC2	general	None
LOC2	price	Negative
LOC2	safety	None
LOC2	transit-location	Positive

Table 1: An example of SentiHood dataset.

- QA-M : “what do you think of the safety of location-1?”
- NLI-M: “location-1-safety”
- QA-B: “The polarity of the aspect safety of location -1 is positive/negative/none”
- NLI-B: “location - 1 - safety - positive/negative/none”

Input Representation

- Sentence $s = \{w_1, \dots, w_m\}$,

Targets $t = \{t_1, \dots, t_k\}$,

Aspect = {general, price, ...}

- Predict sentiment polarity

$y \in \{\text{positive}, \text{negative}, \text{none}\}$

over $\{(t, a) : t \in T, a \in A\}$



Fine tuning and Hyper Parameters

- BERT - base (Transformers block - 12, hidden layer size - 768, parameters - 110M)
- Classification layer
- Softmax layer
- Epochs - 4 , learning rate - 2e-5, batch size 24

Discussion

- Providing more context by adding an auxiliary question
- BERT model has an advantage dealing with sentence pair classification task - supervised masked language model and next sentence prediction task
- The modeling of the question probably also contributed the accuracy in the sentiment classification

Some Other Methods

- Using context from different languages
 - They use the LDA method to find correlation of words from different languages
 - Basically , find correlated words from one language. if you are not able to find the correlation in this specific language, Try to find the correlation in some other language and translate to this language.
- Domain Adaptation using BERT
 - Use the uncased BERT or XLNET , try to train your model after training from BERT with your domain specific data. So that the BIAS of the BERT with your DOMAIN can be overwritten.
 - This also allows BERT to be domain adapted , but also depends on the complexity of the DOMAIN.

How to deal with star ratings?

- Binarization of the star ratings
 - Use regression instead of a binary classifier.
-

Opinion Spamming

- Types of Spam
 - Type 1 (fake reviews)
 - Type 2 (reviews about brands only)
 - Type 3 (non-reviews)
-

Types of Data, Features and Detection

- Three main types of data have been used for review spam detection:
 - Review content
 - Meta-data about the review
 - Product information
-

“Let me look at reviews on one site only...”

Problems?



- Biased views
 - all reviewers on one site may have the same opinion
- Fake reviews/Spam (sites like YellowPages, CitySearch are prone to this)
 - people post good reviews about their own product OR services
 - some posts are plain spams



Coincidence or Fake?



Reviews for a moving company from YellowPages

- # of merchants reviewed by the each of these reviewers → 1
- Review dates close to one another
- All rated 5 star
- Reviewers seem to know exact names of people working in the company and TOO many positive mentions

THE BEST!!!! 11/30/2007 Posted by c_karen ★★★★★

NorthStar did an outstanding job of packing and moving my things. Quite frankly I was expecting some things to be broken. However, to my surprise not one thing was broken and everything went as smooth as could be expected. I had approximately 15,000 lbs. of items to move. I am very impressed with NorthStar and I would not hesitate to utilize them again for my next move. All of the young men who assisted in packing and loading were very hard working and polite

Pros: everything was great

GOOD MOVING 10/11/2007 Posted by joanlee777 ★★★★★

About a month ago, on Sep 12, we hired NorthStar Moving to move our belongings from our house in Van Nuys to the Highway Storage place in Santa Clara. We would like to express our sincere thanks and appreciation for the professional work that was carried out by NorthStar team of workers. In particular, we would like to mention the four NorthStar workers: Roy Ashual, Moshiko Haziza, Guillermo Molise and Roberto Mendoza for their very dedicated service. Besides being good natured and helpful, they worked very well and took good care of our personal effects. We would definitely refer them and NorthStar Moving to any of our friends who are looking for a good moving company.

Great movers 10/08/2007 Posted by shelly_morgan ★★★★★

I wanted to thank the Northstar Moving group for a fabulous job. We hired Northstar Moving on August 4th to move us out of two storage units and where we were staying to our new home in Los Angeles. I had gone through surgery on the 2nd and was in no condition to move around a lot. The Northstar Moving team was great. I slept in while my husband met them at the first pick-up point. Then they came to the 2nd and that is where I met them. When we arrived at the new house they found something for me to sit on and I sat in one place in the garage telling them which room the items went. They were great. They had wonderful personalities. I have never had so much fun moving (even if I was in some pain). Northstar thank you again for the great team and customer service.

Supervised Spam Detection

- Opinion spam detection can be formulated as a classification problem with two classes, fake and non-fake.
 - Due to the fact that there is no labeled training data for learning, **Jindal and Liu (2008)** exploited duplicate reviews.
 - In their study of 5.8 million reviews and 2.14 million reviewers from amazon.com, a large number of duplicate and near-duplicate reviews were found.
-

Four categories to handle duplicates and near duplicates

- Duplicates from the same user-id on the **same** product
 - Duplicates from **different** user-ids on the **same** product
 - Duplicates from the **same** user-id on **different** products
 - Duplicates from **different** user-ids on **different** products
-

Feature engineering for fake reviews

- Review centric features
 - Reviewer centric features
 - Product centric features
-

What is Subjectivity?

- The **linguistic** expression of somebody's **opinions**, **sentiments**, **emotions**.....(private states)
 - private state: state that is not open to objective verification (*Quirk, Greenbaum, Leech, Svartvik (1985). A Comprehensive Grammar of the English Language.*)
 - **Subjectivity analysis** - is the computational study of **affect**, **opinions**, and **sentiments** expressed in text
 - blogs
 - editorials
 - reviews (of products, movies, books, etc.)
 - newspaper articles
-

Example: iPhone



Lab test: Apple gets iPhone 3G right for business An abundance of new features carries iPhone 3G and iPhone 2.0 into the enterprise

By Tom Yager
July 24, 2008

Talkback E-mail Printer Friendly Reprints Text Size A A

InfoWorld

- summary is structured
- everything else is plain text
- mixture of objective and subjective information
- no separation between positives and negatives

With the review iPhone 3G their test and You

Related

New M extra S

AT&T s tetheri

Popula apple,

See A

iPhone deliv

more misses

than hits

iPhone: The \$1,975 iPod

» Back to special report:
Apple launches the iPhone 3G

3G

The Bottom Line

Apple iPhone 3G

Apple, apple.com/iphone

Very Good 8.5

criteria score weight

Extensibility 7 20%

Messaging 8 20%

Networking 9 20%

Usability 9 20%

Multimedia 10 10%

Value

ActiveSync, Assisted GPS (A-GPS), and 1Mbps 3G cell; the iPhone 2.0 software, Apple's new iPhone firmware later will update existing iPhones and iPod Touches to you'll end up with a device that is, except for GPS and 3G, iPhone 3G. The iPod Touch is also upgradable to iPhone 3G.

I've taken to referring to first-gen iPhone and iPhone 3G iPhone, which now identifies a consistently implemented Mac covers all Apple client computers. Wherever I that I'm making specific reference to Apple's new hand

Second time's the charm

Apple has turned iPhone into a mobile platform that can and enterprise users. I make that recommendation without testing of the iPhone 3G against Apple's claims. Those some time. It's my opinion that final judgment about this can't be rendered until you've trusted your digital identity.

Clearly, I haven't had time to carry it that far, but the iPhone software meet the expectations set by Apple, and Apple produced a mobile device and platform that hold their own E-Series, RIM BlackBerry, and Windows Mobile 6. In its innovation, the iPhone 3G exceeds the competition.

makes it hard to imagine competitors closing the gap.

Review on InfoWorld -
tech news site

s for the device, a e 3G and the new among other things, an a cellular browser

Product summary

Good:

Apple iPhone has a stunning display, a sleek and an innovative multitouch user interface. Safari browser makes for a superb Web surfing experience, and it offers easy-to-use apps. As an iPhone shines.

Bad:

Example: iPhone



Lab test: Apple gets iPhone 3G right for business

An abundance of new features carries iPhone 3G and iPhone 2.0 into the enterprise

By Tom Yager
July 24, 2008

Talkback E-mail Printer Friendly Reprints Text Size A A

With the iPhone 3G's banner opening weekend and newsstands looking like a rack of brochures for the device, a review of the iPhone 3G at this point might be pro forma, except for one thing: Much of the iPhone 3G and the new iPhone 2.0 software remains an enigma to professionals and enterprises, users set apart by, among other things, their tendency to use punctuation in their e-mail. These users demand more from a handset than a cellular browser and YouTube.

Related Stories
New MacBook Air: now with extra SSD goodness
AT&T says iPhone 3G tethering coming 'soon'
Popular Tags
apple, iphone-3g

See Also
iPhone delivers more misses than hits
iPhone: The \$1,975 iPod

**» Back to special report:
Apple launches the iPhone 3G**

The Bottom Line
Apple iPhone 3G
Apple, apple.com/iphone

Very Good 8.5
criteria score weight
Extensibility 7 20%
Messaging 8 20%
Networking 9 20%
Usability 9 20%
Multimedia 10 10%

Clearly, I haven't had time to carry it that far, but the iPhone 3G's software meets the expectations set by Apple, and Apple has produced a mobile device and platform that hold their own against the E-Series, RIM BlackBerry, and Windows Mobile 6. In its innovation, the iPhone 3G exceeds the competition.

makes it hard to imagine competitors closing the gap.
www.infoworld.com/article/0807/0707TECH01.html

Review on InfoWorld - tech news site

Product summary

The good:

The Apple iPhone has a stunning display, a sleek design, and an innovative multitouch user interface. Its Safari browser makes for a superb Web surfing experience, and it offers easy-to-use apps. As an iPod, it shines.

The bad:

The Apple iPhone has variable call quality and lacks some basic features found in many cell phones, including stereo Bluetooth support and 3G compatibility. Integrated memory is stingy for an iPod, and you have to sync the iPhone to manage music content.

The bottom line:

Despite some important missing features, a slow data network, and call quality that doesn't always deliver, the Apple iPhone sets a new benchmark for an integrated cell phone and MP3 player.

Specifications:

OS provided: Apple MacOS X; Band / mode: GSM 850/900/1800/1900 (Quadband); Wireless connectivity: IEEE 802.11b, IEEE 802.11g, Bluetooth 2.0 EDR; [See full specs](#)

[See all products in the Apple iPhone series](#)

CNET editors' review

Reviewed by: Kent Gerber

Edited by: Lindsey Turner

Reviewed on: 06/30/2008

Updated on: 07/11/2008

CNET review

Review posted on a tech blog

iP...
Pub...



See my NEW iPhone 3g review

Let me start off by saying that while I'm a fan of Apple's success and products, I'm not one of those people that blindly apologizes for their products no matter what. I'll be the first to say that something works or it doesn't. My friends and many of you come to me all the time because they want my HONEST assessment. So I wanted a couple of days with the iPhone to really take it through its paces and see if this new phone is what it's hyped up to be. You must also understand that there isn't a smartphone out there that I think is perfect. As a matter of fact before the iPhone there were basically 4 smartphone OS's, Palm, Blackberry, Symbian and Windows Mobile. I stuck with Palm because it was the lesser of the 4 evils or the one that sucks least. Palm has a UI (user interface) that is... zero innovation. However, there are thousands of OS. Blackberry doesn't have a touch screen or tap... pad/thumb wheel. Also the Blackberry's I considered etc.) Symbian looked very promising, but I was how sloooooow it was and that there were very few up on him just last week right in front of me.

Subjectivity Analysis on iPhone Reviews



Individual's Perspective

- Highlight of what is good and bad about iPhone
 - Ex. Tech blog may contain mixture of information
 - Combination of good and bad from the different sites (*tech blog, InfoWorld and CNET*)
 - Complementing information
 - Contrasting opinions
- Ex.

CNET: *The iPhone lacks some basic features*

Tech Blog: *The iPhone has a complete set of features*

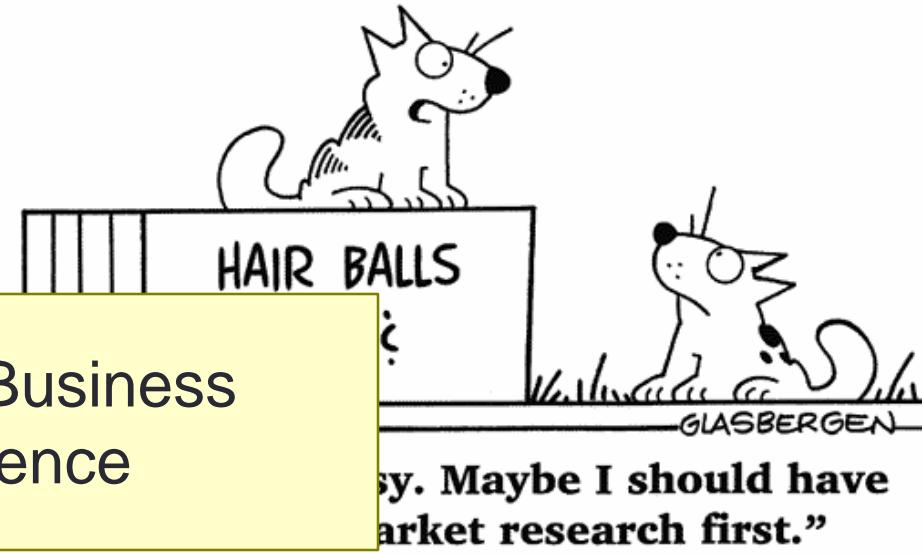
Subjectivity Analysis on iPhone Reviews



Business' Perspective

- Apple: What do consumers think about iPhone?
 - Do they like it?
 - What do they dislike?
 - What are the major complaints?
 - What features should we add?
- Apple's competitor:
 - What are iPhone's weaknesses?
 - How can we compete with them?
 - Do people like

Known as Business
Intelligence



Google Product Search



HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$89 online, \$100 nearby 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews

1 star 2 3 4 stars 5 stars

What people are saying

ease of use



"This was very easy to setup to four computers."

value



"Appreciate good quality at a fair price."

setup



"Overall pretty easy setup."

customer service



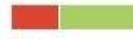
"I DO like honest tech support people."

size



"Pretty Paper weight."

mode



"Photos were fair on the high quality mode."

colors



"Full color prints came out with great quality."

Bing Shopping

HP Officejet 6500A E710N Multifunction Printer

[Product summary](#) [Find best price](#) **Customer reviews** [Specifications](#) [Related items](#)



\$121.53 - \$242.39 (14 stores)

Compare

Average rating (144)

(55)

(54)

(10)

(6)

(23)

(0)

Most mentioned

Performance

(57)

Show reviews by source

Best Buy (140)

Ease of Use

(43)

CNET (5)

Print Speed

(39)

Amazon.com (3)

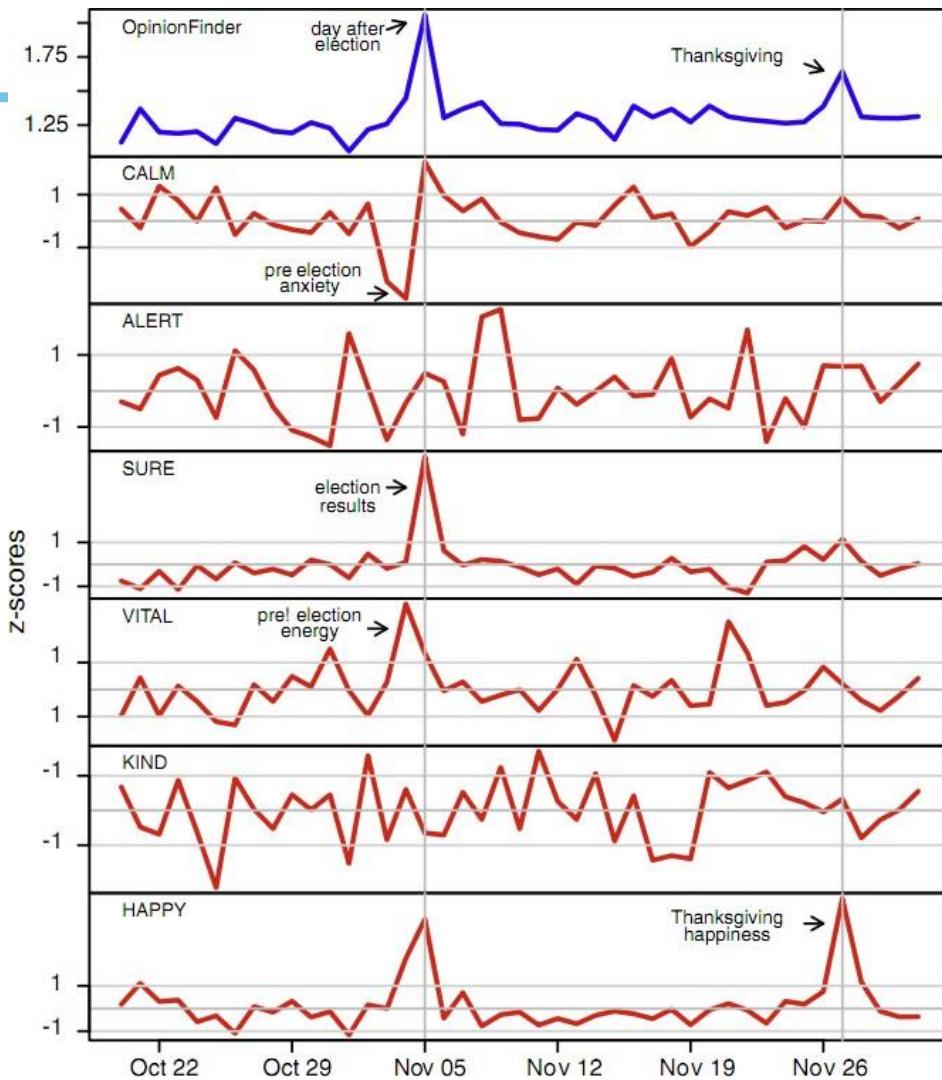
Connectivity

(31)

More ▾

Twitter sentiment:

Johan Bollen, Huina Mao, Xiaojun Zeng.
[2011. Twitter mood predicts the stock market,](#)
 Journal of Computational Science 2:1, 1-8. 10.1016/j.jocs.2010.12.007.



Target Sentiment on Twitter

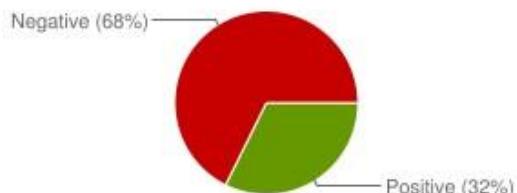
Type in a word and we'll highlight the good and the bad

- [Twitter Sentiment App](#)
- Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision

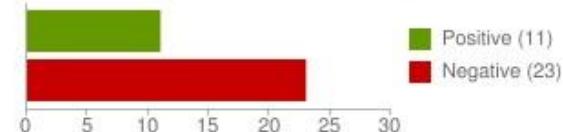
"united airlines" [Save this search](#)

Sentiment analysis for "united airlines"

Sentiment by Percent



Sentiment by Count



jjacobsen: OMG... Could @United airlines have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human.
Posted 2 hours ago

12345clumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?
Posted 2 hours ago

EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination. <http://t.co/Z9QloAif>
Posted 2 hours ago

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more, but cell phones off now!
Posted 4 hours ago

Application Areas Summary

- Businesses and organizations: interested in opinions
 - product and service benchmarking
 - market intelligence
 - survey on a topic
 - Individuals: interested in other's opinions when
 - Purchasing a product
 - Using a service
 - Tracking political topics
 - Other decision making tasks
 - Ads placements: Placing ads in user-generated content
 - Place an ad when one praises a product
 - Place an ad from a competitor if one criticizes a product
 - Opinion search: providing general search for opinions
 - Text-driven forecasting: insights about other areas from text
-

Application of sentiment analysis

- Business and organization
 - Market research



- Customer service
- Ads placements-Social media
 - Place an ad if one praises the product
 - Place an ad from competitor if one criticizes the product
- Individual
 - Make decisions to purchase products or to use services
 - Find public opinions about political candidates and issues.

References

- https://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00049
- <https://alphabold.com/sentiment-analysis-the-lexicon-based-approach/>
- <https://web.eecs.umich.edu/~mihalcea/papers/banea.lrec08.pdf>
- <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>
- https://www.youtube.com/watch?v=OEUIzQawd1s&feature=emb_logo
- [VADER Sentiment Analysis: A Complete Guide, Algo Trading and More \(quantinsti.com\)](https://quantinsti.com/vader-sentiment-analysis-a-complete-guide-algo-trading-and-more/)
- <https://www.kaggle.com/datasets/naveedhn/amazon-product-review-spam-and-non-spam>
- <https://tech.hindustantimes.com/tech/news/amazon-fake-review-scam-discovered-affects-nearly-200-000-users-here-s-how-it-worked-71620616179506.html>
- <https://www.kaggle.com/datasets/rtatman/sentiment-lexicons-for-81-languages>

References

- <https://www.kaggle.com/datasets/rtatman/sentiment-lexicons-for-81-languages>
 - <https://github.com/aesuli/SentiWordNet>
 - <https://github.com/cjhutto/vaderSentiment>
 - <https://towardsdatascience.com/tensorflow-sarcasm-detection-in-20-mins-b549311b9e91>
 - <https://www.geeksforgeeks.org/sentiment-analysis-of-hindi-text-python/?ref=rp>
 - <https://www.youtube.com/playlist?list=PL83F70cPvROYoMqibhzo3zB88dcOUj07Q>
 - <https://www.youtube.com/watch?v=szcZpgOEdXs&t=90s>
 - <https://www.youtube.com/watch?v=q8sTicXK4Fg>
 - <https://github.com/ScalaConsultants/Aspect-Based-Sentiment-Analysis>
 - <https://github.com/declare-lab/multimodal-deep-learning/tree/main>
-

References

- <https://colab.research.google.com/drive/1DQcywfg7IXrsXbNTeauKruhRadnoizME#scrollTo=uJcg5S9ujSLb>
- https://drive.google.com/drive/folders/1ya2UGUuTjE_YmNv9kw6F3vP-Cd-Up7H7
- <https://drive.google.com/drive/folders/1TK9k41RT8Nf3lhzerNWHpEqWztsk2gAP>
- https://colab.research.google.com/drive/1Pa3M_NtsBiHCQ_1A2fudLfEQEyClwDv0
- <https://www.youtube.com/watch?v=q8sTicXK4Fg>
- <https://github.com/ScalaConsultants/Aspect-Based-Sentiment-Analysis>
- <https://www.youtube.com/watch?v=bkq-pA5Avcq>
- <https://towardsdatascience.com/sentiment-analysis-in-10-minutes-with-bert-and-hugging-face-294e8a04b671>
- <https://www.analyticsvidhya.com/blog/2021/12/fine-tune-bert-model-for-sentiment-analysis-in-google-colab/>



Thank you