Natural Language Processing Applications

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

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

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

innovate achieve lead

Statistical Machine translation

- Introduction
- Approaches
- Parallel Corpora
- Word Alignment
- Language Models
- Translation Models
- IBM Models
- MT Evaluation
- Bleu Score

Machine Translation

innovate achieve lead

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!)*

1 minute video showing 1954 MT: https://youtu.be/K-HfpsHPmvw

- *Systems were mostly rule-based, using a bilingual dictionary to map Russian words to their English counterparts*

Rule Based Systems

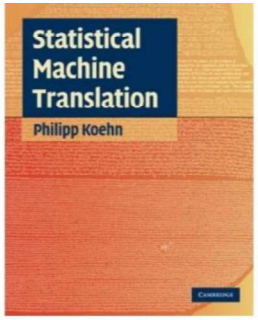
- Rule-based systems
  - build dictionaries
  - write transformation rules
  - refine, refine, refine
- Météo system for weather forecasts (1976)
- Systran (1968), Logos and Metal (1980s)

```
"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.

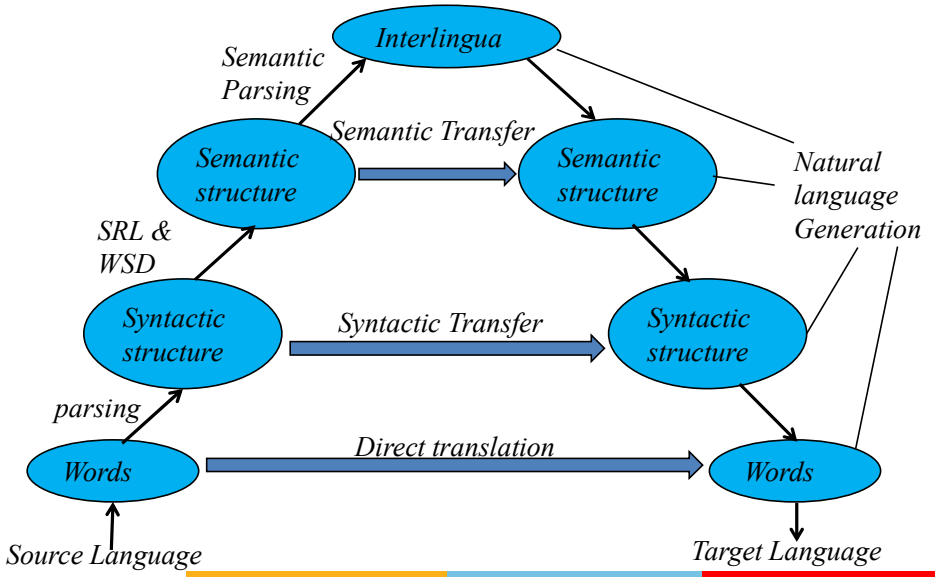
Since you brought it up, we have been working on it.

- 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



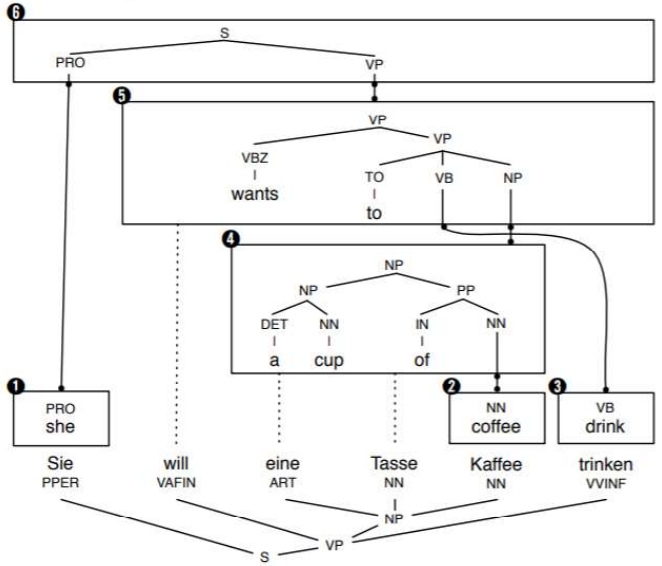
Direct Transfer

- 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 dar: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 → Adj Nom ⇒ NP → Nom ADJ
  - English to Japanese:
    - VP → V NP ⇒ VP → NP V
    - PP → P NP ⇒ PP → NP P

Syntax-based Translation



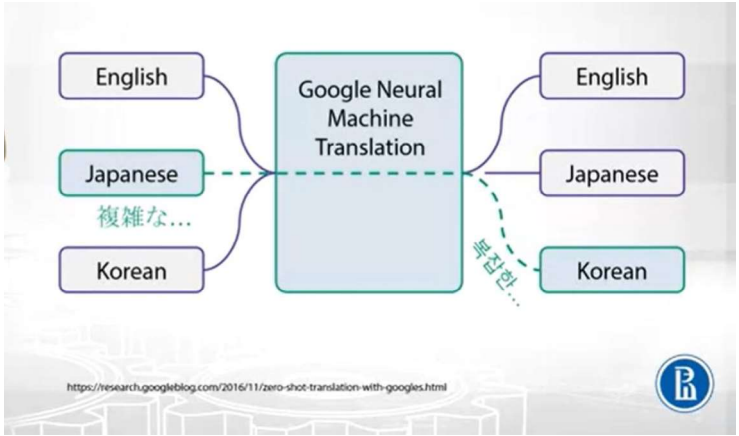
Semantic Transfer

- 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 → V PP[+benefactor] ⇒ VP → PP[+benefactor] V

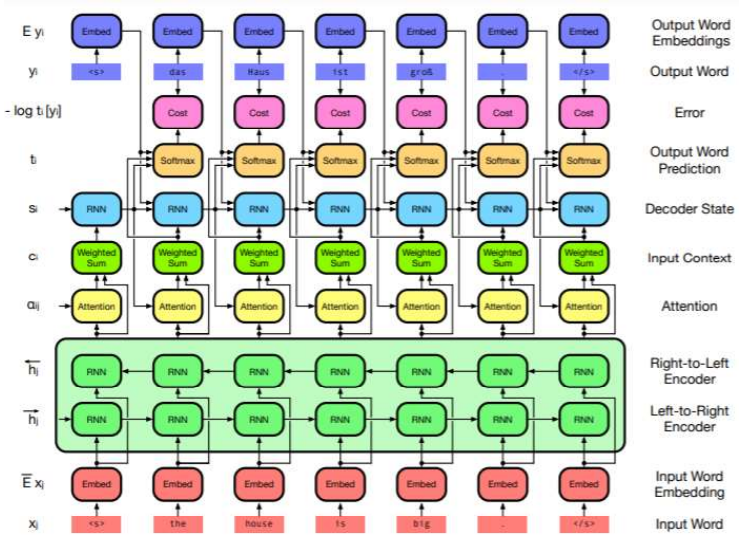
Semantic Transfer

- 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*

$$\text{argmax}_y P(y|x)$$

- *Use Bayes Rule to break this down into two components to be learnt separately:*

$$= \text{argmax}_y P(x|y)P(y)$$

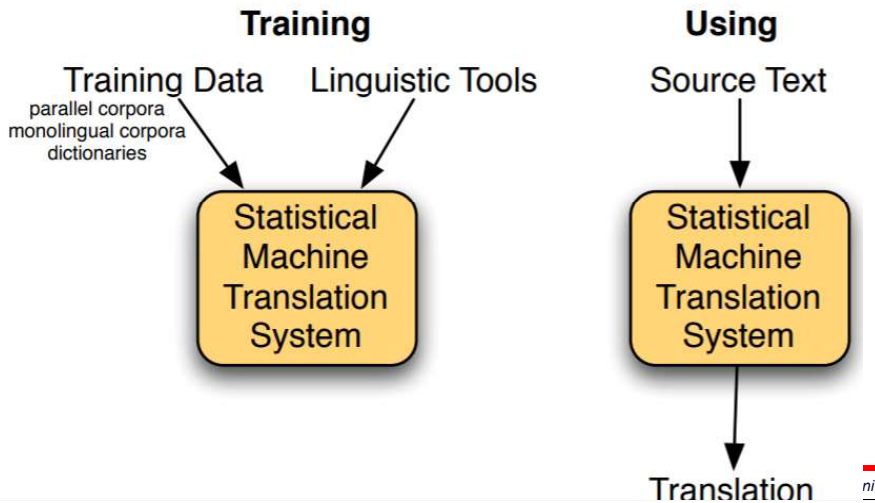
Translation Model

Models how words and phrases should be translated (fidelity). Learnt from parallel data.

Language Model

Models how to write good English (fluency). Learnt from monolingual data.

Statistical Machine Translation



Parallel Corpora

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	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 *Haus*	Count
house	8,000
building	1,600
home	200
household	150
shell	50

Maximum likelihood estimation

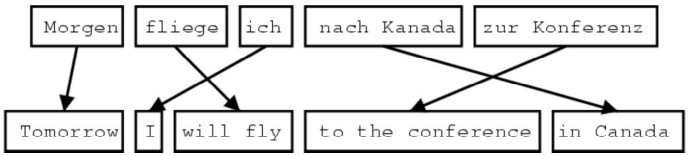
$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \text{house}, \\ 0.16 & \text{if } e = \text{building}, \\ 0.02 & \text{if } e = \text{home}, \\ 0.015 & \text{if } e = \text{household}, \\ 0.005 & \text{if } e = \text{shell}. \end{cases}$$

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

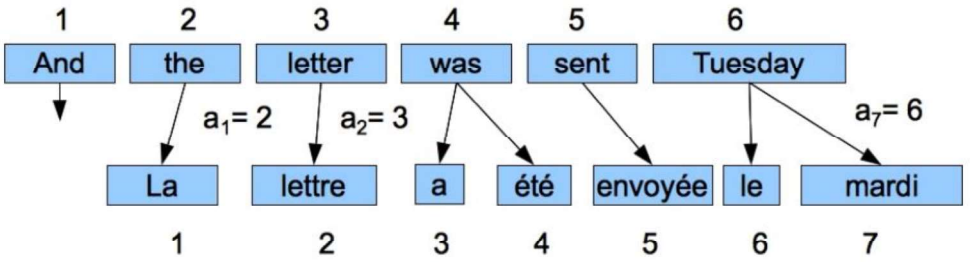
- *We learn* $P(x, a|y)$ *as a combination of many factors, including:*
  - *Probability of particular words aligning (also depends on position in sent)*
  - *Probability of particular words having particular fertility (number of corresponding words)*
  - *etc.*
- *Alignments a are latent variables: They aren't explicitly specified in the data!*
  - *Require the use of special learning aglos (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

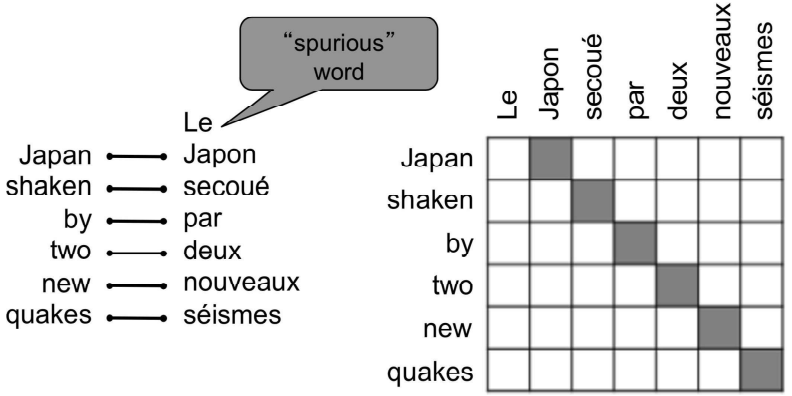
- Shows mapping between words in one language and the other.



What is alignment?

Alignment is the correspondence between particular words in the translated sentence pair.

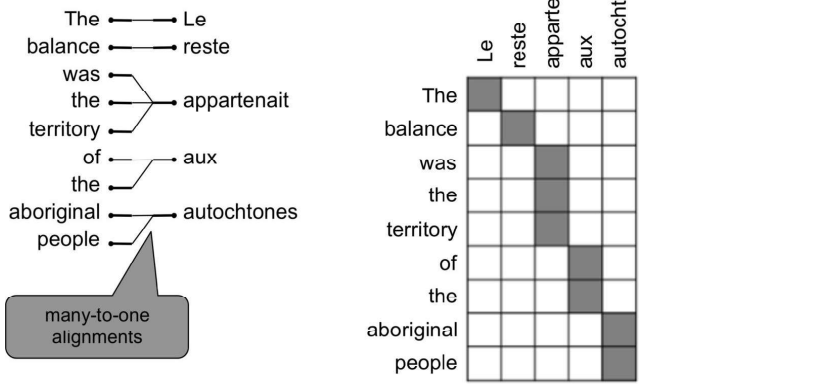
- *Typological differences between languages lead to complicated alignments!*
- *Note: Some words have no counterpart*



Examples from: *"The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993.*
http://www.aclweb.org/anthology/J93-2003

Alignment is complex

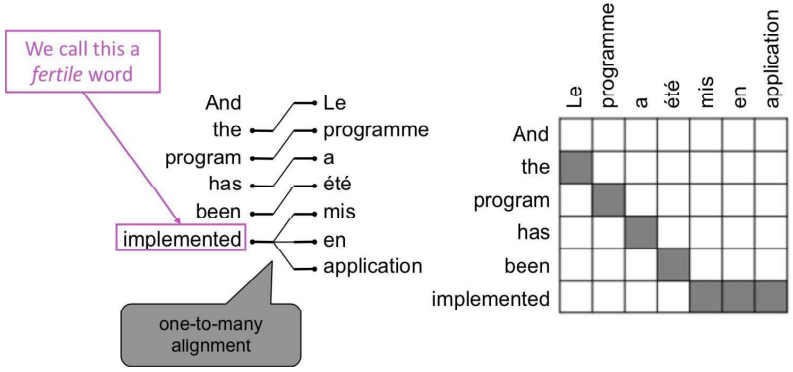
Alignment can be many-to-one



Examples from: *"The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993.*
http://www.aclweb.org/anthology/J93-2003

Alignment is complex

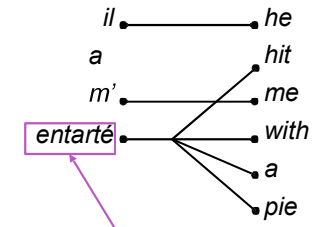
Alignment can be one-to-many

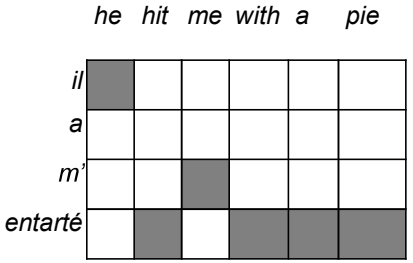


Examples from: *"The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993.*
http://www.aclweb.org/anthology/J93-2003

Alignment is complex

Some words are very fertile!

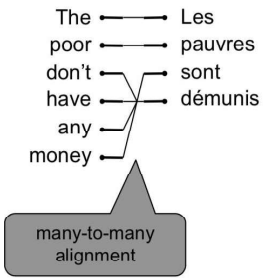


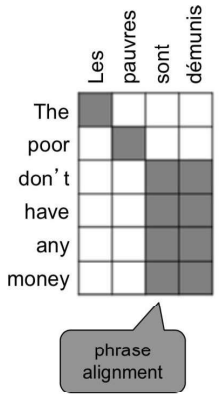


This word has no single-word equivalent in English

Alignment is complex

Alignment can be many-to-many (phrase-level)





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 $f=f_1, f_2, \ldots f_m$ to an English sentence $\hat{e} = e_1, e_2, \ldots e_I$ that maximizes $P(e \mid f)$

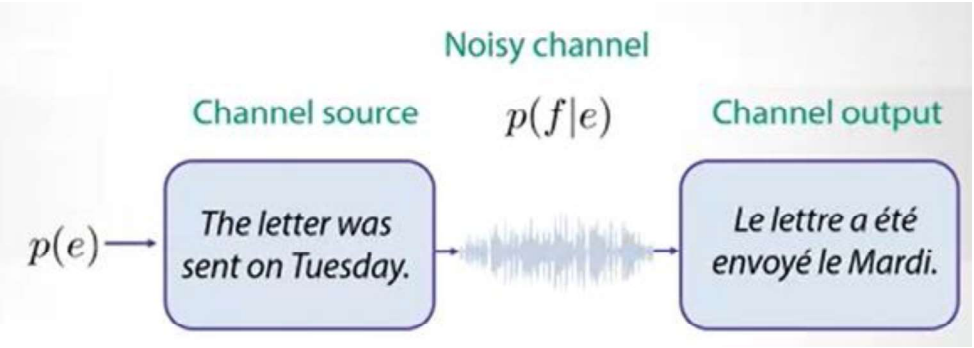
Bayesian Analysis of Noisy Channel

$$\hat{e} = \text{argmax } P(e \mid f)$$
$$= \text{argmax } P(f \mid e)P(e) / P(f)$$
$$= \text{argmax } P(f \mid e)P(e)$$

Translation Model *Language Model*

A decoder determines the most probable translation $\hat{e}$ given f

Noisy Channel: Translation Model



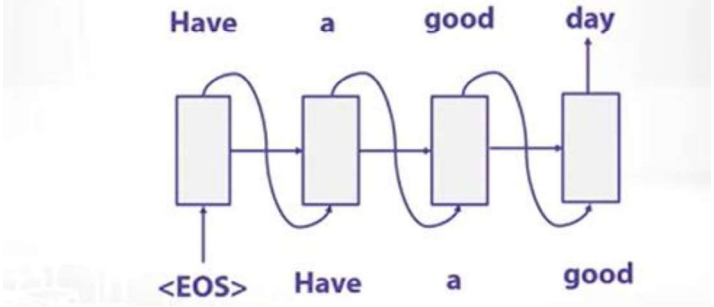
Language Model

- 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

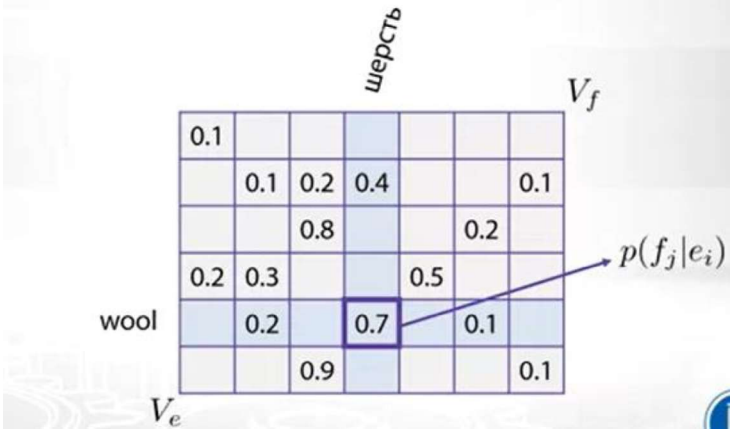
$$p(f|e) = p(f_1, f_2, \ldots f_J|e_1, e_2, \ldots e_I)$$

f (Foreign): Крику много, а шерсти мало.

e (English): Great cry and little wool.

Translation Model

We could learn translation probabilities for separate words:



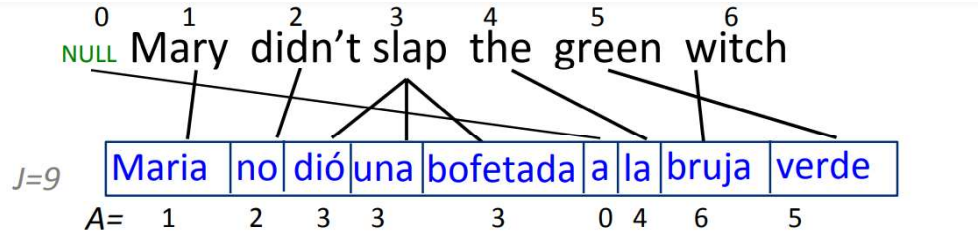
IBM Model 1

- 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 $E=e_1, e_2, \ldots e_I$

- Choose J, the number of words in F: $F=f_1, f_2, \ldots f_J$
- Choose a 1-to-many alignment $A=a_1, a_2, \ldots a_J$
- 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: $F=f_1, f_2, \ldots f_J$
2. Choose a 1-to-many alignment $A=a_1, a_2, \ldots a_J$
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:

$$P(F \mid E, A) = \prod_{j=1}^{J} t(f_j, e_{a_j})$$

- 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:

$$P(A \mid E) = \frac{\varepsilon}{(I+1)^J}$$

- The probability of an alignment given the English sentence.

IBM Model 1

$$P(A \mid E) = \frac{\varepsilon}{(I+1)^J} \qquad P(F \mid E, A) = \prod_{j=1}^{J} t(f_j, e_{a_j})$$

The probability of generating F through a particular alignment:

$$P(F, A \mid E) = \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j, e_{a_j})$$

To get P(*F* | *E*), we sum over all alignments:

$$P(F \mid E) = \sum_{A} P(F, A \mid E) = \sum_{A} \frac{\varepsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j, e_{a_j})$$

IBM Model 1

- Goal is to find the most probable alignment given a parameterized model.

$$\begin{aligned}\hat{A} &= \underset{A}{\operatorname{argmax}}\, P(F, A \mid E) \\ &= \underset{A}{\operatorname{argmax}}\, \frac{P(J \mid E)}{(I+1)^J} \prod_{j=1}^{J} t(f_j, e_{a_j}) \\ &= \underset{A}{\operatorname{argmax}} \prod_{j=1}^{J} t(f_j, e_{a_j})\end{aligned}$$

Since translation choice for each position *j* is independent, the product is maximized by maximizing each term:

$$a_j = \underset{0 \le i \le I}{\operatorname{argmax}}\, t(f_j, e_i) \quad 1 \le j \le J$$

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

$$P(A,F \mid E) = \prod_{j=1}^{J} t(f_j \mid e_{a_j})$$

(ignoring a constant here)

- Normalize to get probability of an alignment:

$$P(A \mid E,F) = \frac{P(A,F \mid E)}{\sum_A P(A,F \mid E)} = \frac{\prod_{j=1}^{J} t(f_j \mid e_{a_j})}{\sum_A \prod_{j=1}^{J} t(f_j \mid e_{a_j})}$$

35

Sample EM Trace for Alignment
(IBM Model 1 with no NULL Generation)

Training Corpus

green house / *casa verde*

the house / *la casa*

Translation Probabilities

	verde	*casa*	*la*
green	1/3	1/3	1/3
house	1/3	1/3	1/3
the	1/3	1/3	1/3

Assume uniform initial probabilities

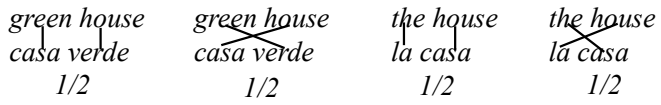
Compute Alignment Probabilities P(a, f | e)

green house / *casa verde*	*green house* / *casa verde*	*the house* / *la casa*	*the house* / *la casa*
1/3 X 1/3 = 1/9	*1/3 X 1/3 = 1/9*	*1/3 X 1/3 = 1/9*	*1/3 X 1/3 = 1/9*

Normalize to get P(a | f, e)

$\frac{1/9}{2/9} = \frac{1}{2}$ $\quad \frac{1/9}{2/9} = \frac{1}{2}$ $\quad \frac{1/9}{2/9} = \frac{1}{2}$ $\quad \frac{1/9}{2/9} = \frac{1}{2}$

Example cont.



Compute weighted translation counts

	verde	casa	la
green	1/2	1/2	0
house	1/2	1/2 + 1/2	1/2
the	0	1/2	1/2

Normalize rows to sum to one to estimate P(f | e)

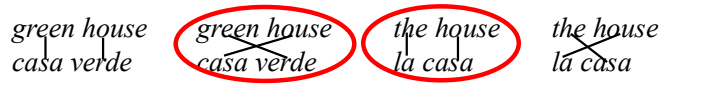
	verde	casa	la
green	1/2	1/2	0
house	1/4	1/2	1/4
the	0	1/2	1/2

Example cont.

Translation Probabilities

	verde	casa	la
green	1/2	1/2	0
house	1/4	1/2	1/4
the	0	1/2	1/2

Recompute Alignment Probabilities P(a, f | e)



Normalize to get P(a | f, e)

$\frac{1/8}{3/8}=\frac{1}{3}$ $\frac{1/4}{3/8}=\frac{2}{3}$ $\frac{1/4}{3/8}=\frac{2}{3}$ $\frac{1/8}{3/8}=\frac{1}{3}$

Continue EM iterations until translation parameters converge

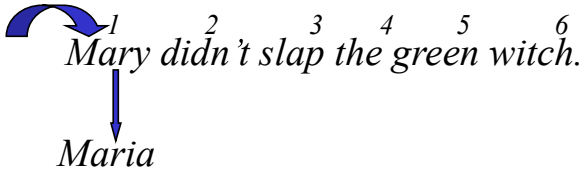
HMM-Based Word Alignment

- 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.

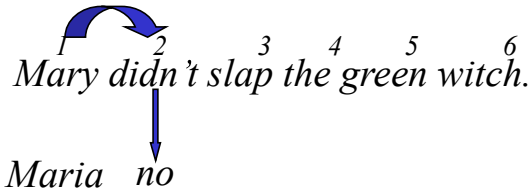
HMM Model

- Assumes the hidden state is the specific word occurrence e_i in ***e*** currently being translated (i.e. there are *I* 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.

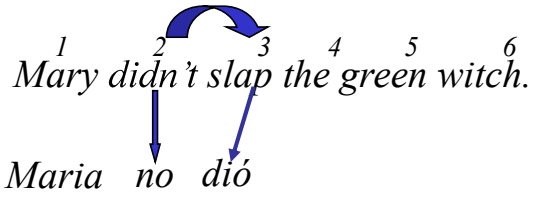
Sample HMM Generation



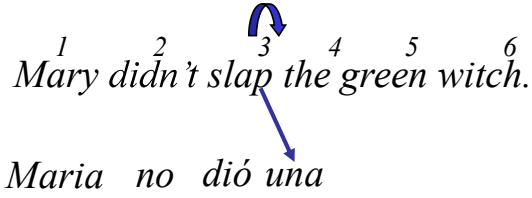
Sample HMM Generation



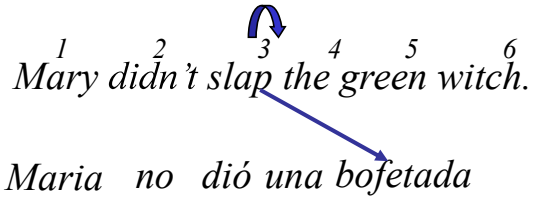
Sample HMM Generation



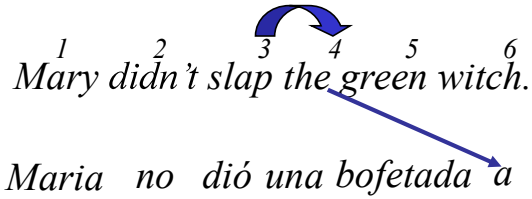
Sample HMM Generation



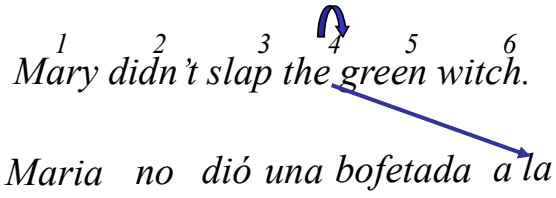
Sample HMM Generation



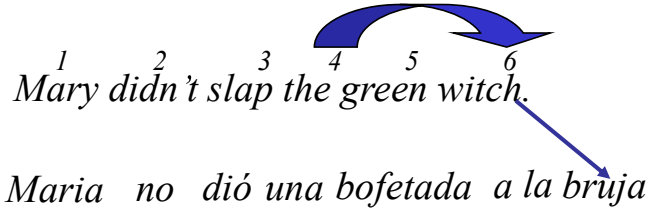
Sample HMM Generation



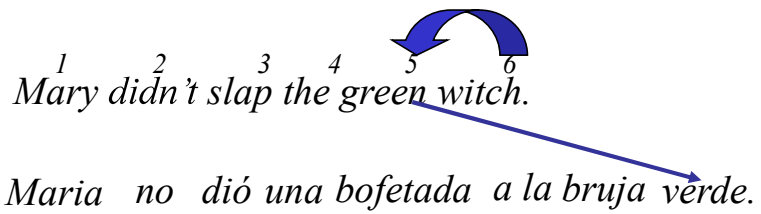
Sample HMM Generation



Sample HMM Generation



Sample HMM Generation



Sample HMM Generation

1 2 3 4 5 6
Mary didn't slap the green witch.
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**: $b_j(f_i)=P(f_i \mid e_j)$ 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:

$$T_{best} = \underset{\text{T} \in \text{Target}}{\text{argmax}}\ \text{faithfulne ss}(T, S)\ \text{fluency}(T)$$

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 ofen 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

$$\text{Precision} = \frac{\text{No. of candidate translation words occuring in any reference translation}}{\text{Total no. of words in the candidate translation}}$$

Candidate 1: the the the the the the the.
Candidate 2: the cat is mat 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: *MIN(R(w), C(w))* where MIN is the minimum of the two values.
Step 4: Bleu Score= Clip count / Total

Bleu- Example 1

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
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
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
MIN(C(but), R(but))=MIN(1, 1)=1
MIN(C(love), R(love))=MIN(3, 2)=2
MIN(C(other), R(other))=MIN(1, 0)=0
MIN(C(friend), R(friend))=MIN(1,0)=0
MIN(C(for), R(for))=MIN(1, 2)=1
MIN(C(yourself), R(yourself))=MIN(1,1)=1
Total clip count=5
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.

$$p_n = \frac{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count}\ (\text{n-gram})} \qquad p = \sqrt[N]{\prod_{n=1}^{N} p_n}$$

Cand 1: $p = \sqrt[2]{\frac{5}{6}\frac{1}{5}} = 0.408$

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.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

BLEU Score

Final BLEU Score: BLEU = $BP \times p$

Cand 1: Mary no slap the witch green.

Best Ref: Mary did not slap the green witch.

$$c = 6, \quad r = 7, \quad BP = e^{(1-7/6)} = 0.846$$

$$BLEU = 0.846 \times 0.408 = 0.345$$

Cand 2: Mary did not give a smack to a green witch.

Best Ref: Mary did not smack the green witch.

$$c = 10, \quad r = 7, \quad BP = 1$$

$$BLEU = 1 \times 0.558 = 0.558$$

BLEU Score Issues

- BLEU has been shown to correlate with human evaluation when comparing outputs from different SMT systems.
- However, it is 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=AIpXjFwVdIE
Dataset: http://www.manythings.org/anki/

Natural Language Processing Applications

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

BITS *Pilani*
Pilani Campus

BITS *Pilani*
Pilani Campus

Session 10: Neural Machine translation
Date – 18th February 2024
Time – 1.40 pm to 3.40 pm

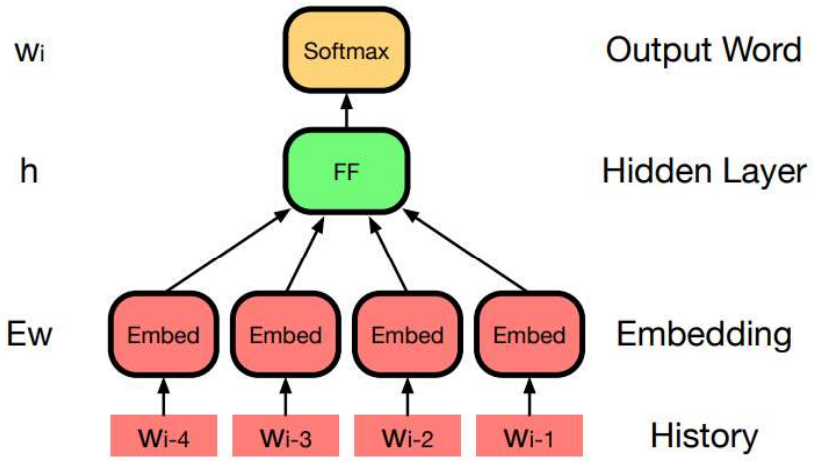
These slides are prepared by the instructor, with grateful acknowledgement of Prof. **Philipp Koehn** , Prof. Raymond J. Mooney, Prof. Jurafsky, Abigail See, Matthew Lamm and many others who made their course materials freely available online.

Agenda

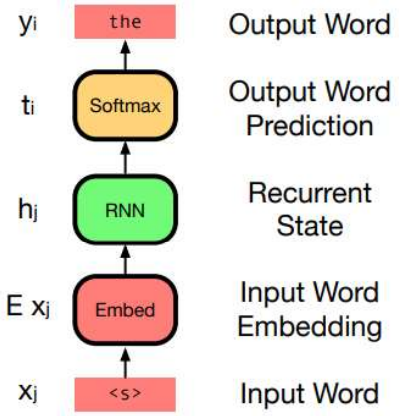
Neural Machine translation

- 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

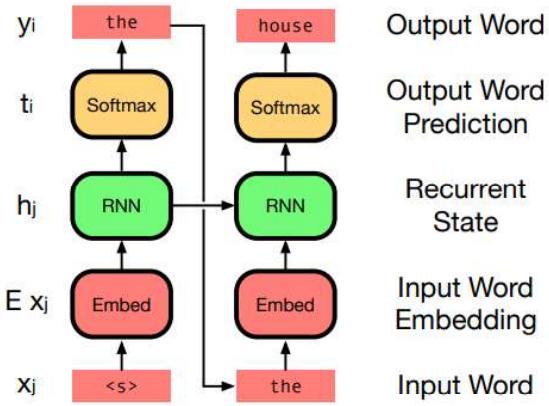


Recurrent Neural Language Model



Predict the first word of a sentence

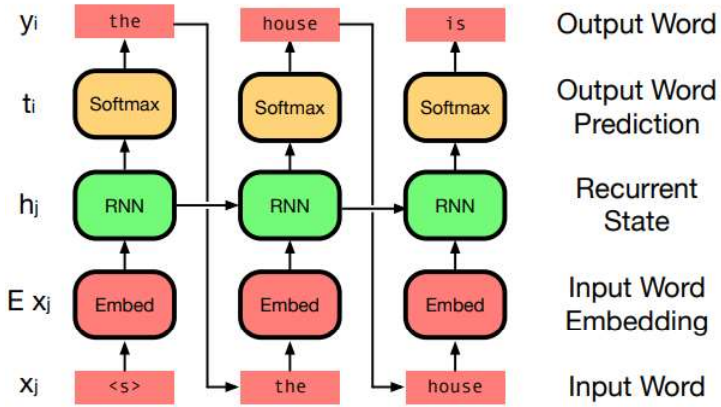
Recurrent Neural Language Model



Predict the second word of a sentence

Re-use hidden state from first word prediction

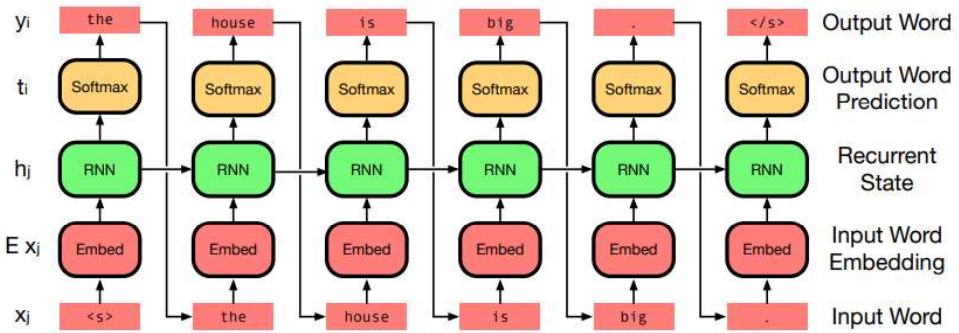
Recurrent Neural Language Model



Predict the third word of a sentence

... and so on

Recurrent Neural Language Model



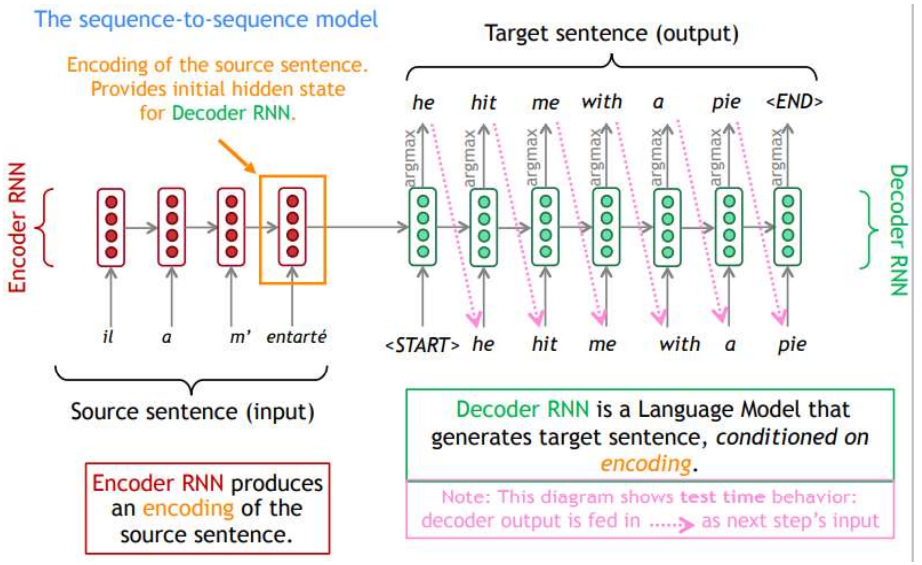
What is Neural Machine Translation?

- 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

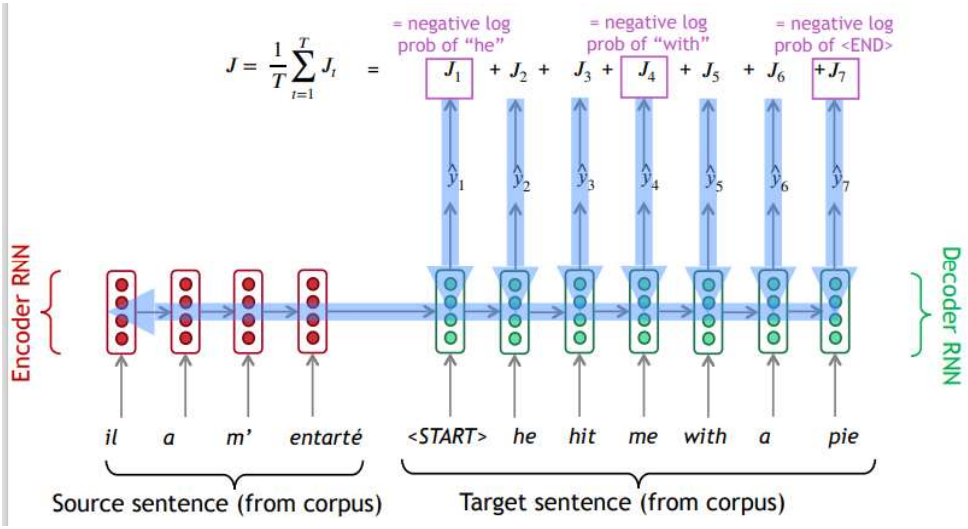
NN Machine translation?

- 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)

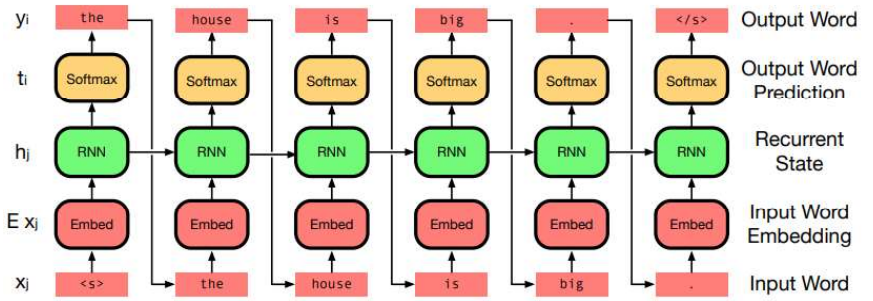


Training NMT



Neural translation model

Input Encoding



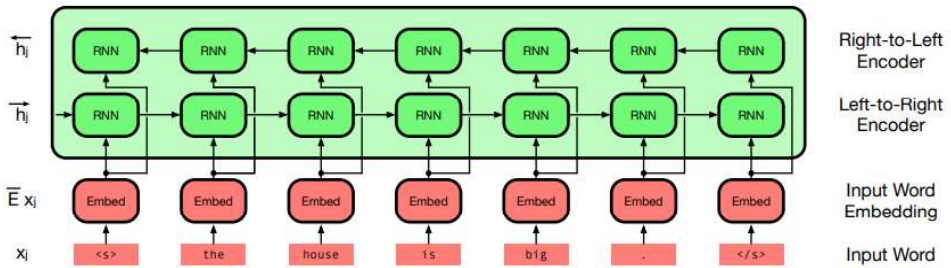
- Inspiration: recurrent neural network language model on the input side

Hidden Language Model States

- This gives us the hidden states

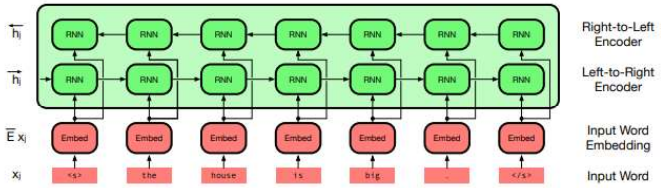
- These encode left context for each word
- Same process in reverse: right context for each word

Input Encoder-Bidirectional RNN



- Input encoder: concatenate bidirectional RNN states
- Each word representation includes full left and right sentence context

Encoder Math



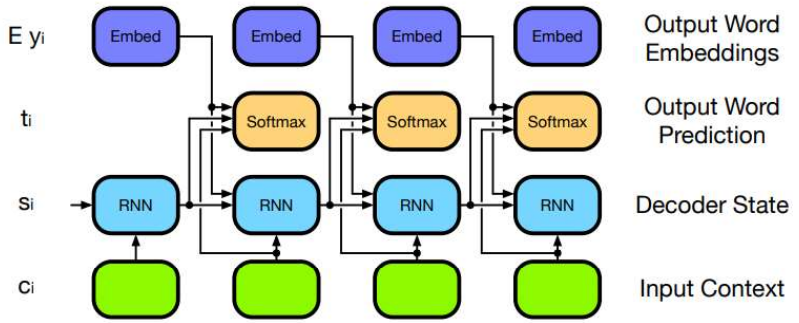
- Input is sequence of words x_j, mapped into embedding space $\bar{E}\, x_j$
- Bidirectional recurrent neural networks

$$\overleftarrow{h_j} = f(\overleftarrow{h_{j+1}}, \bar{E}\, x_j)$$
$$\overrightarrow{h_j} = f(\overrightarrow{h_{j-1}}, \bar{E}\, x_j)$$

- Various choices for the function $f()$: feed-forward layer, GRU, LSTM, ...

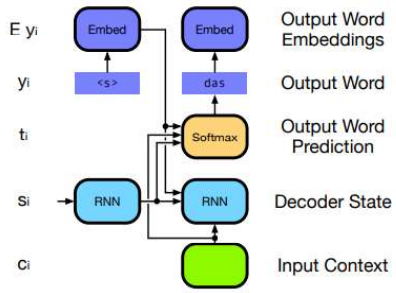
Decoder

- 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

$$s_i = f(s_{i-1},\ Ey_{-1}, c_i)$$

- 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)

$$t_i = W(Us_{i-1} + VEy_{i-1} + Cc_i)$$

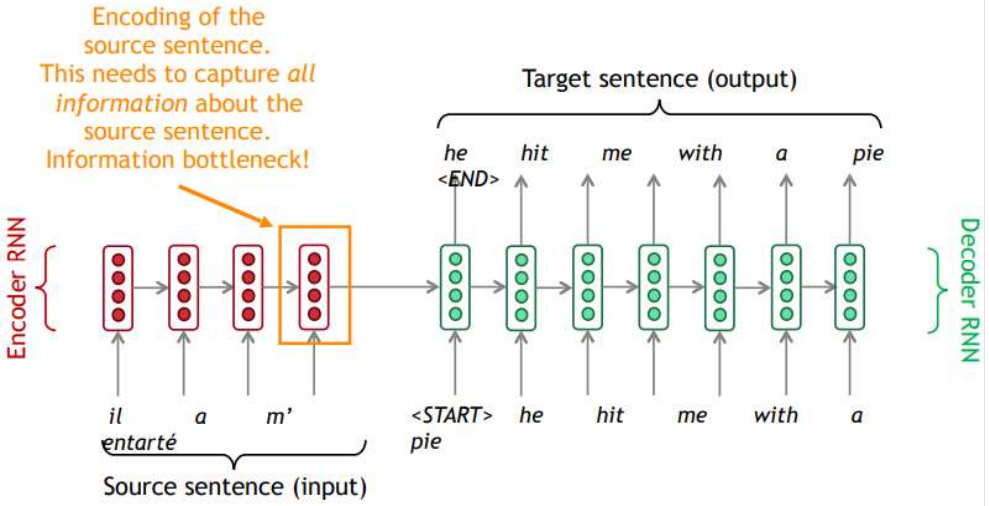
then finding the highest value in vector t_i

- If we normalize t_i, we can view it as a probability distribution over words
- Ey_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

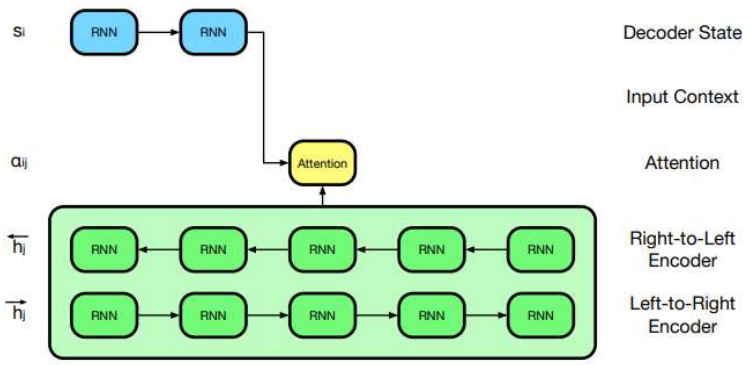
Bottleneck



Attention

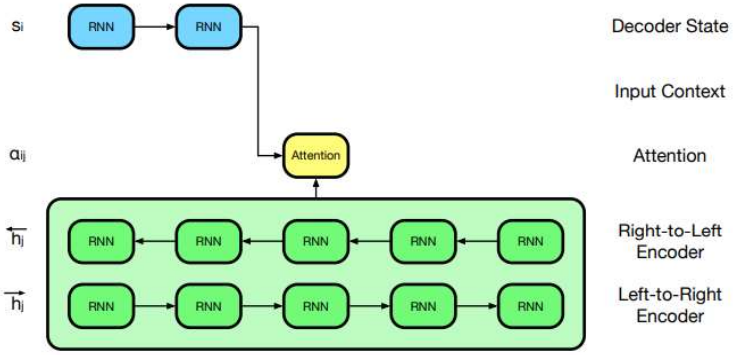
- Attention provides a solution to the bottleneck problem.
- Core idea: on each step of the decoder, use direct connection to the encoder to focus on a particular part of the source sequence
- First we will show via diagram (no equations), then we will show with equations

Attention



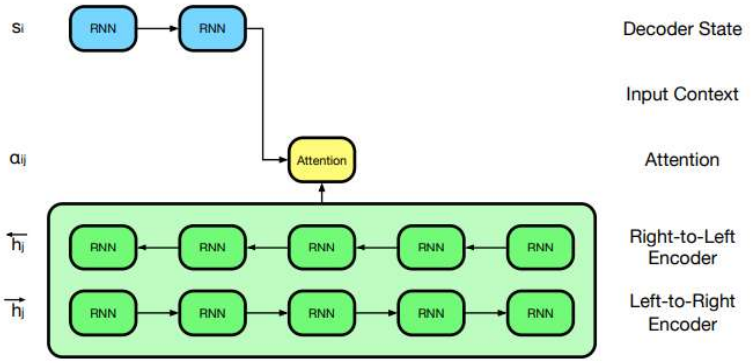
- Given what we have generated so far (decoder hidden state)
- ... which words in the input should we pay attention to (encoder states)?

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 with a feed-forward neural network layer)

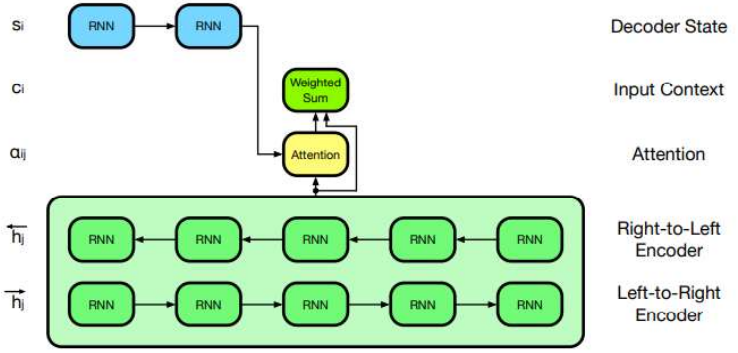
Attention



- Normalize attention (softmax)

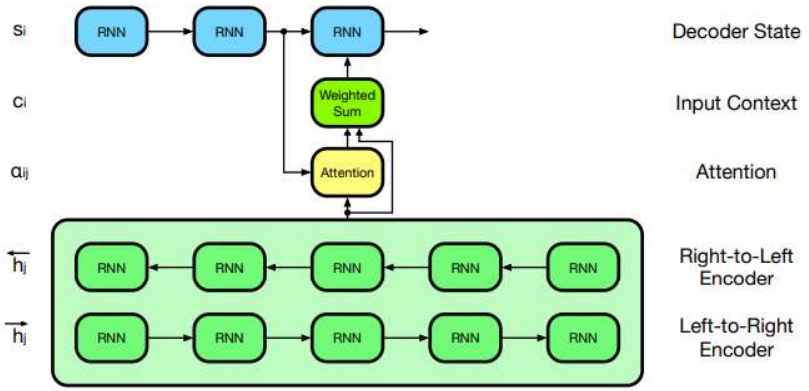
$$\alpha_{ij} = \frac{\exp(a(s_{i-1}, h_j))}{\sum_k \exp(a(s_{i-1}, h_k))}$$

Attention



- Relevant input context: weigh input words according to attention: $c_i = \sum_j \alpha_{ij} h_j$

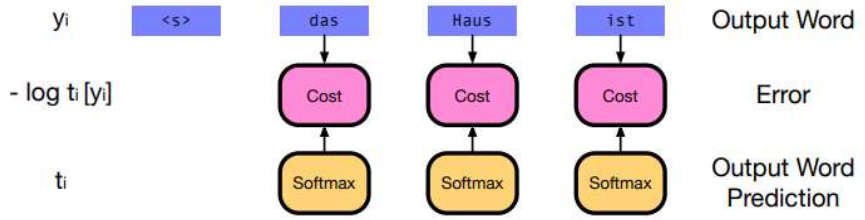
Attention



Use context to predict next hidden state and output word

Training Phase

Comparing predicted word with actual

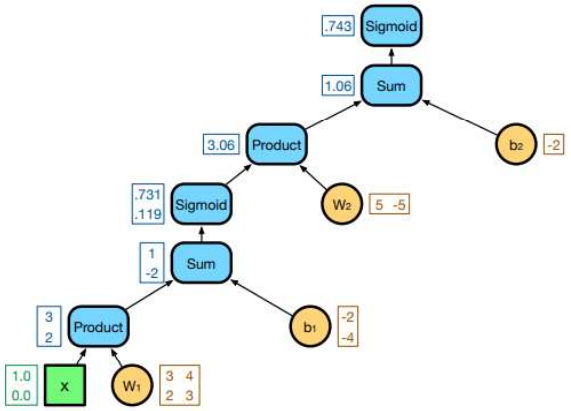


Current model gives some probability $t_i[y_i]$ to correct word y_i

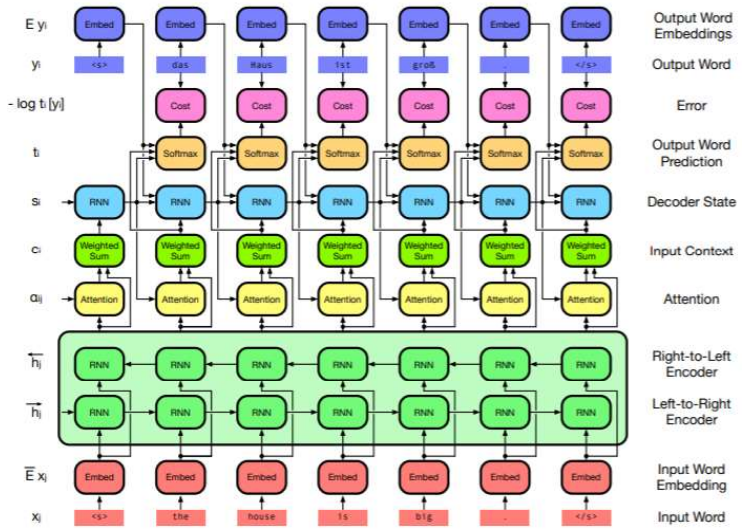
We turn this into an error by computing cross-entropy: $-\log t_i[y_i]$

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

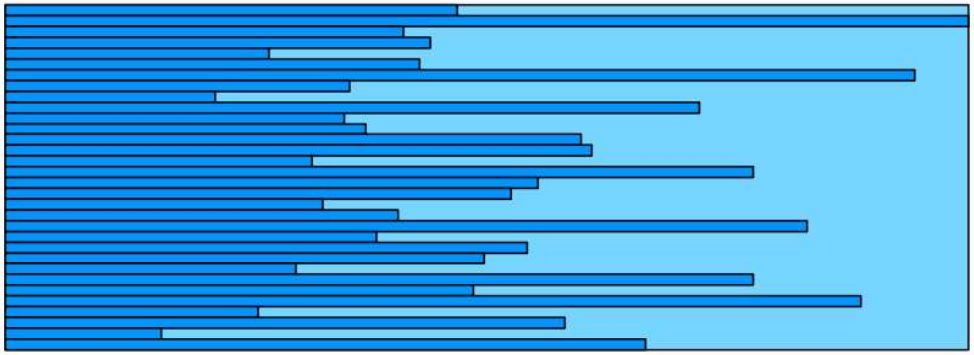
- Already large degree of parallelism
- most computations on vectors, matrices
- efficient implementations for CPU and GPU •

Further parallelism by batching

- • - processing several sentence pairs at once
- • Typical batch sizes 50-100 sentence pairs

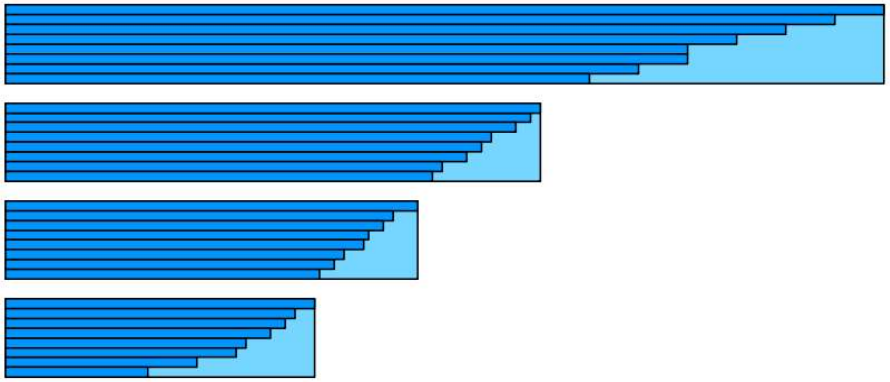
Batching

- 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

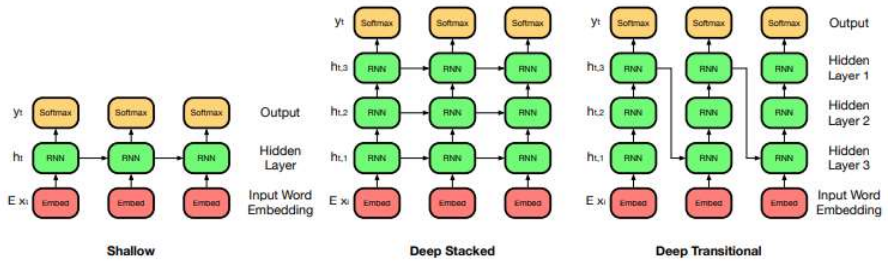


Organization of training

- 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

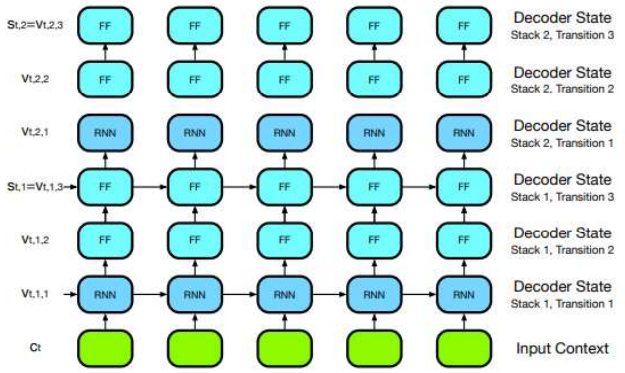
- Encoder and decoder are recurrent neural networks
- We can add additional layers for each step
- Recall shallow and deep language models



- Adding residual connections (short-cuts through deep layers) help

Deep Decoder

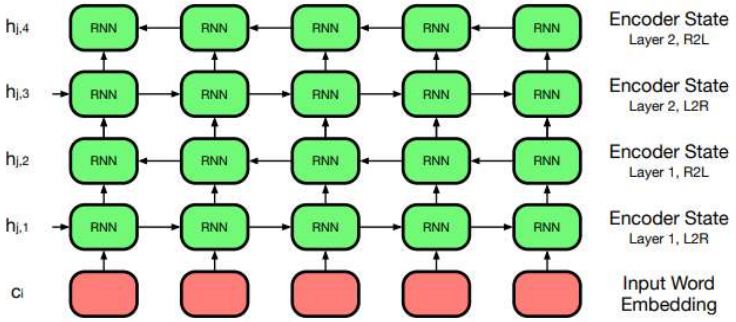
- Two ways of adding layers
  - deep transitions: several layers on path to output
  - deeply stacking recurrent neural networks
- Why not both?



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



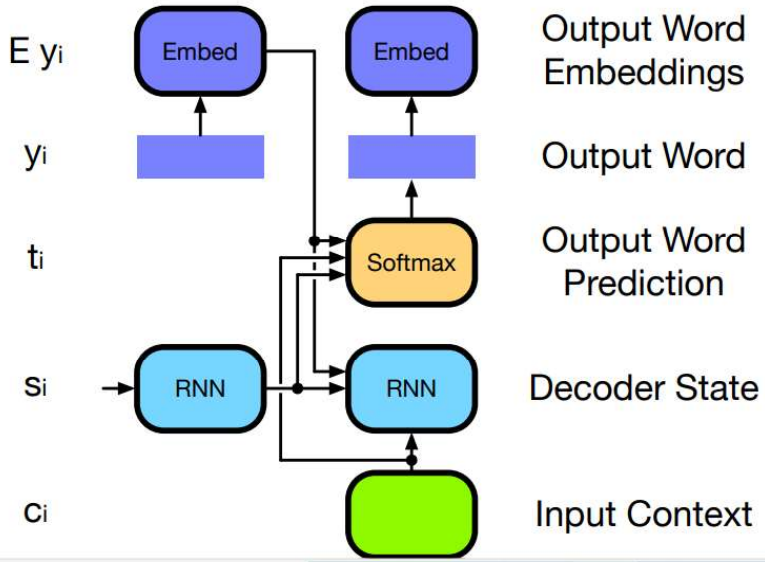
Inference

- Given a trained model

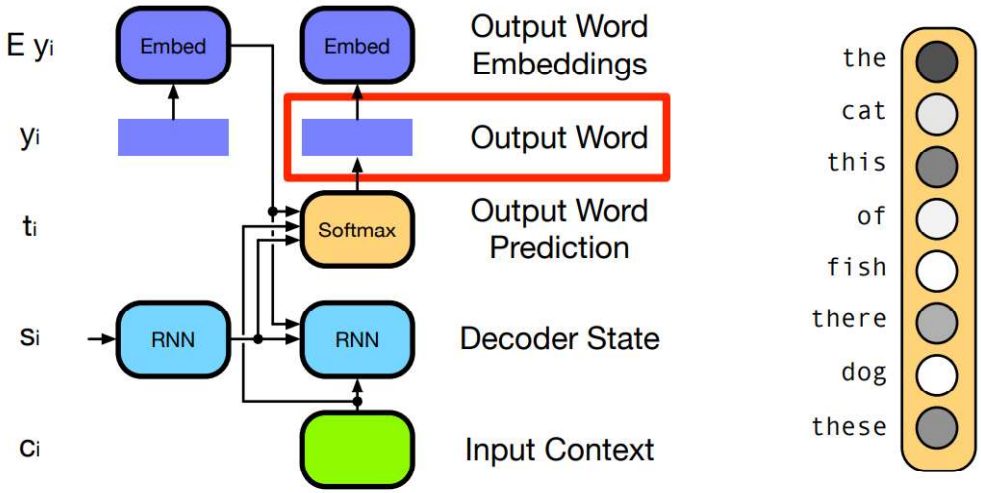
... we now want to translate test sentences

- We only need execute the ”forward” step in the computation graph

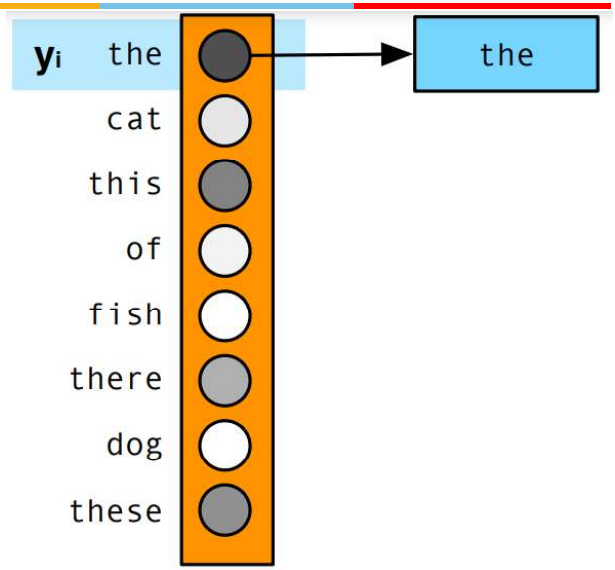
Word Prediction



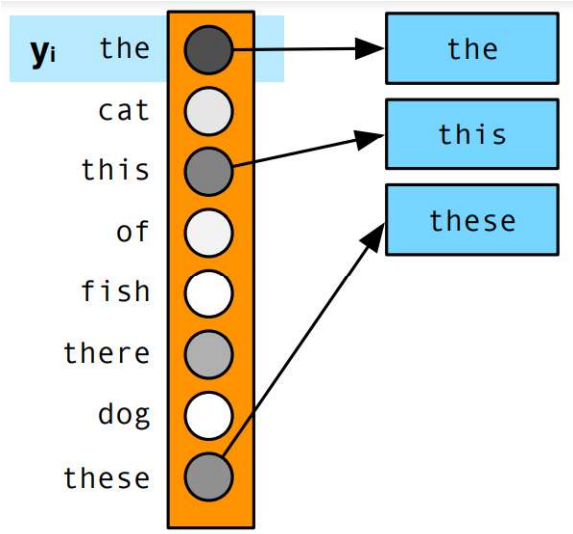
Selected word



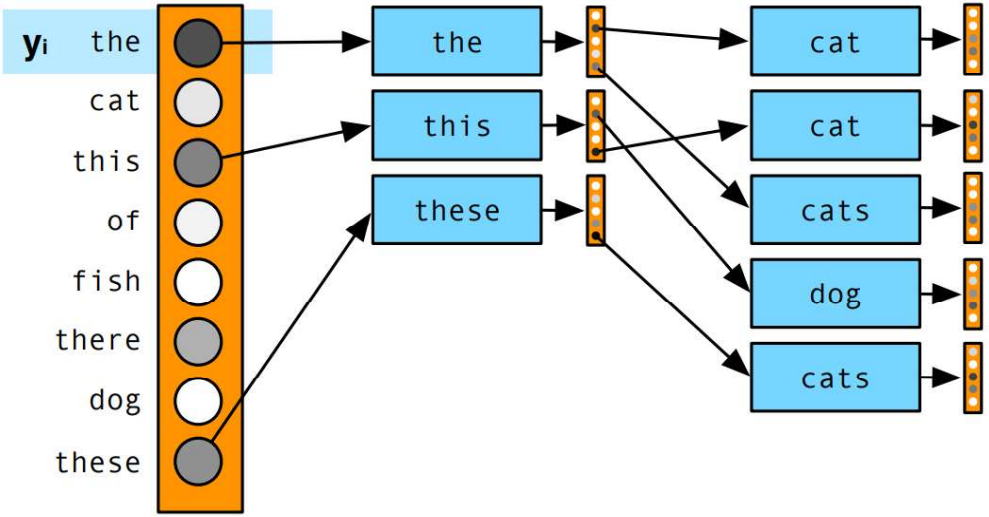
Select the best word



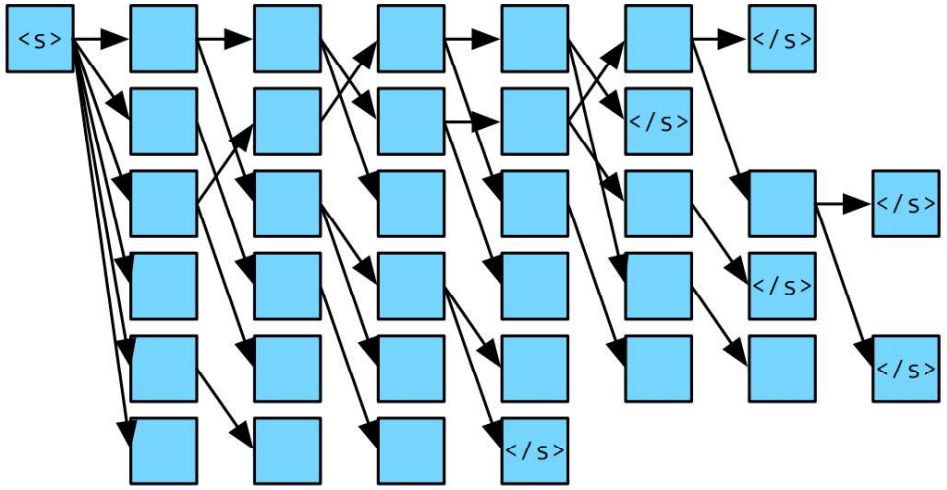
Select second and third best



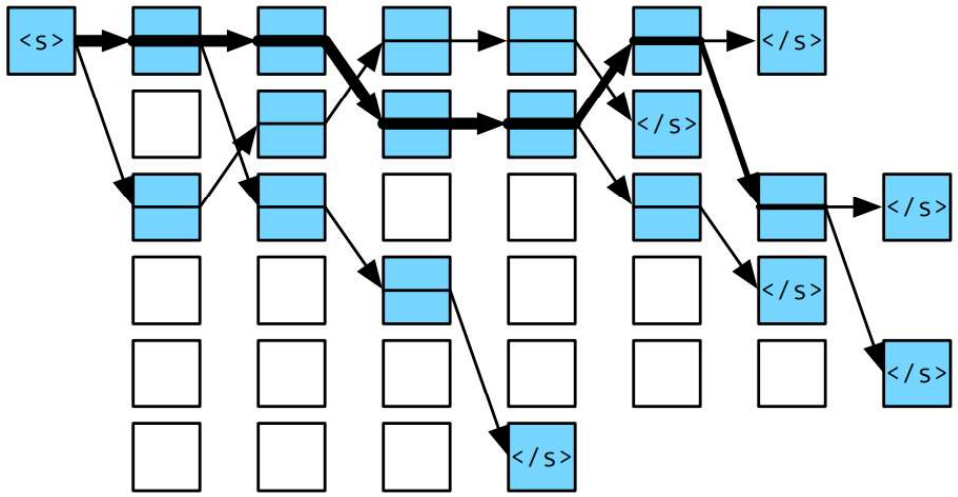
Select best continuation



Beam Search



Best path

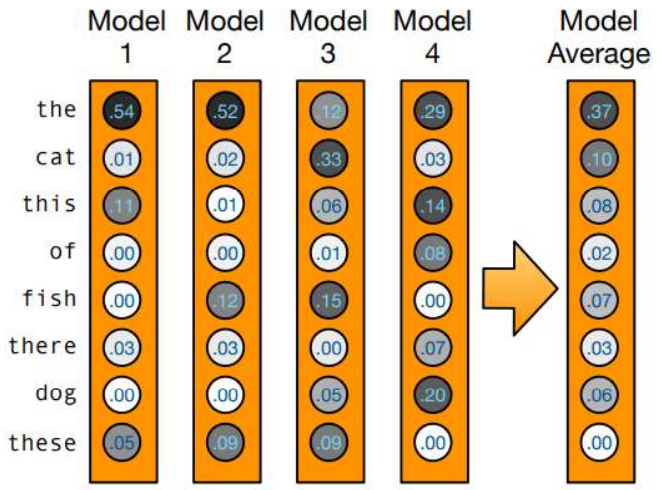


Output word predictions

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%)	
to	(95.5%)	*about* (1.2%), *for* (1.1%), *in* (1.0%), *of* (0.3%), *around* (0.1%), ...
keep	(69.8%)	*maintain* (4.5%), *hold* (4.4%), *be* (4.2%), *have* (1.1%), *make* (1.0%), ...
his	(86.2%)	*its* (2.1%), *statements* (1.5%), *what* (1.0%), *out* (0.6%), *the* (0.6%), ...
statements	(91.9%)	*testimony* (1.5%), *messages* (0.7%), *comments* (0.6%), ...
vague	(96.2%)	*v@@* (1.2%), *in* (0.6%), *ambiguous* (0.3%), ...
enough	(98.9%)	*and* (0.2%), ...
so	(51.1%)	*,* (44.3%), *to* (1.2%), *in* (0.6%), *and* (0.5%), *just* (0.2%), *that* (0.2%), ...
they	(55.2%)	*that* (35.3%), *it* (2.5%), *can* (1.6%), *you* (0.8%), *we* (0.4%), *to* (0.3%), ...
can	(93.2%)	*may* (2.7%), *could* (1.6%), *are* (0.8%), *will* (0.6%), *might* (0.5%), ...
be	(98.4%)	*have* (0.3%), *interpret* (0.2%), *get* (0.2%), ...
interpreted	(99.1%)	*interpre@@* (0.1%), *constru@@* (0.1%), ...
in	(96.5%)	*on* (0.9%), *differently* (0.5%), *as* (0.3%), *to* (0.2%), *for* (0.2%), *by* (0.1%), ...
different	(41.5%)	*a* (25.2%), *various* (22.7%), *several* (3.6%), *ways* (2.4%), *some* (1.7%), ...
ways	(99.3%)	*way* (0.2%), *manner* (0.2%), ...
.	(99.2%)	*</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)

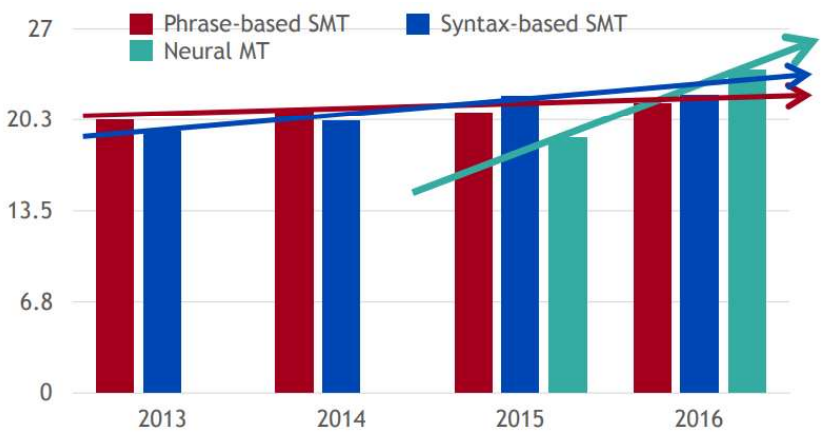
  the ← cat ← is ← in ← the ← bag ← .

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 ⇒ use both left and right context during translation
- Only possible once full sentence produced → 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!

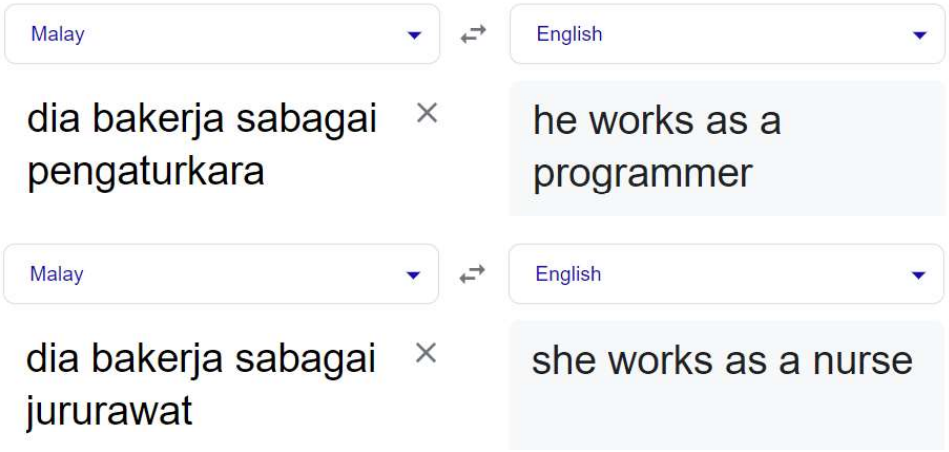
- Many difficulties remain:
- Out-of-vocabulary words
- Domain mismatch between train and test data
- Maintaining context over longer text
- Low-resource language pairs

Is Machine Translation solved?

- Using common sense is still hard
- Idioms are difficult to translate

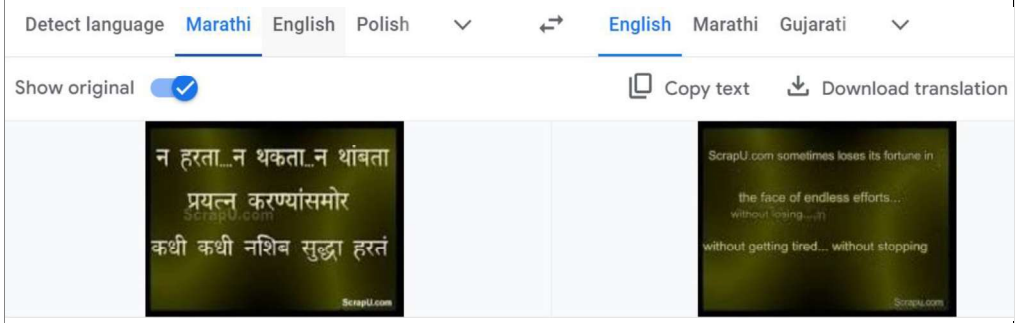
Is Machine Translation solved?

NMT picks up biases in training data



Is Machine Translation solved?

Uninterpretable systems do strange things



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=AIpXjFwVdIE

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.

innovate achieve lead

Agenda

Indic Machine translation

- Introduction
- Recurrent Neural Translation Models
- Encoder Decoder Translation Models
- Neural translation model with attention
- Training Neural Models
- Deeper Models
- Demo

innovate achieve lead

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 !!***

innovate achieve lead

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?

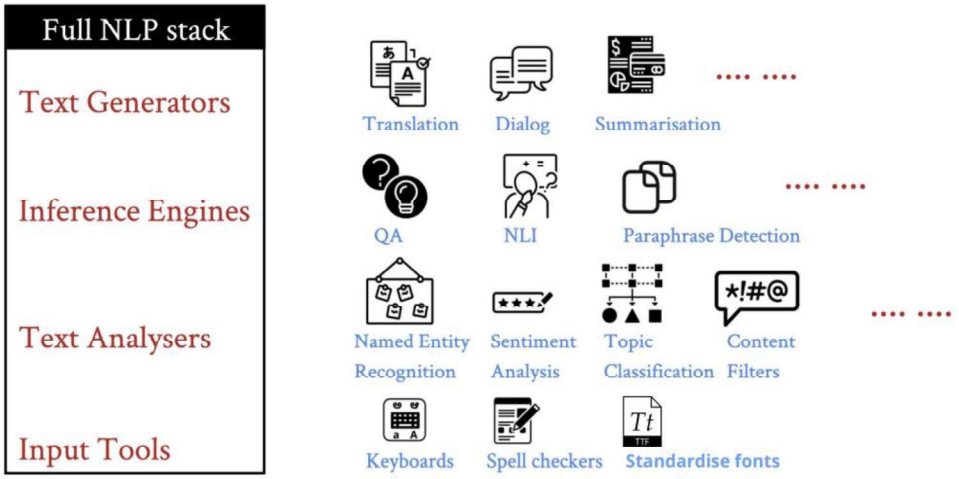
Automatic conversion of text/speech from Indian language to another languae

Be the change you want to see in the world

वह परिवर्तन बनो जो संसार में देखना चाहते हो

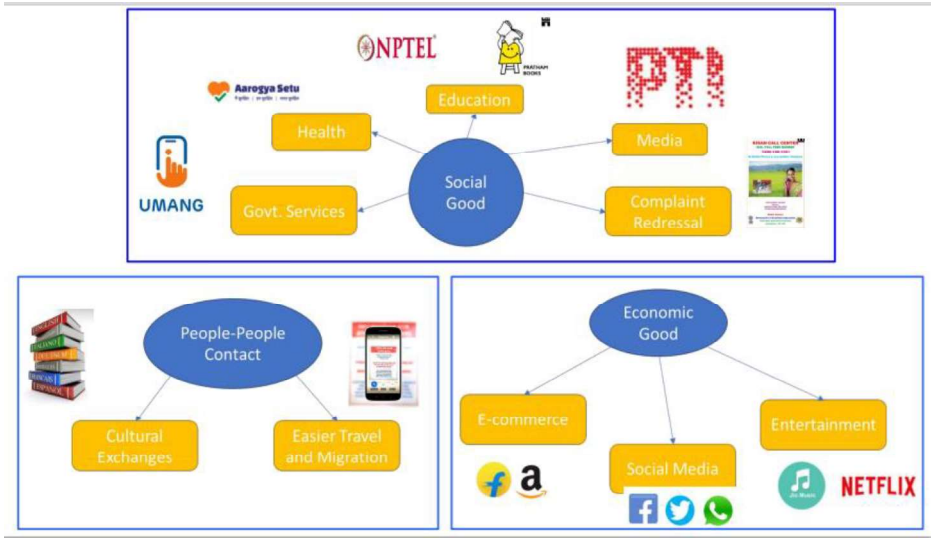
bing Translator Google Translate SDL TRADOS SYSTRAN beyond language

NLP for 22 constitutional languages



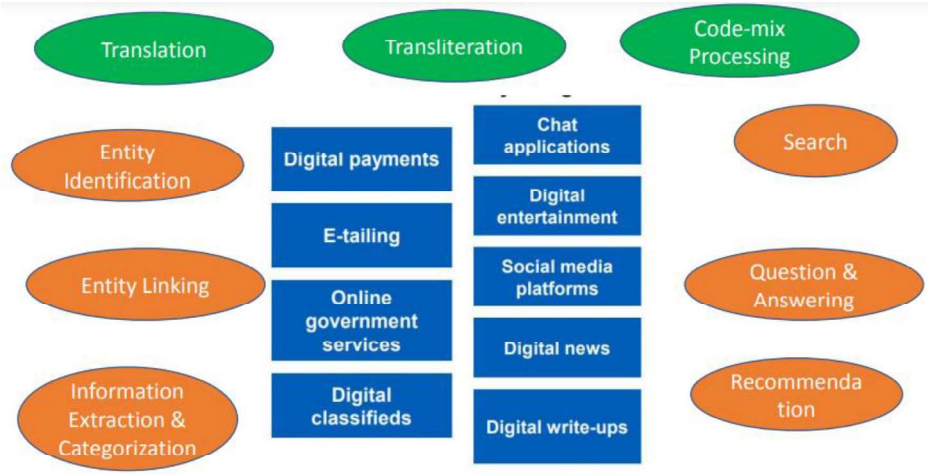
Ref: AI4Bharat Tutorial at ICON 2021, Dec 2021

Applications- Indian Languages



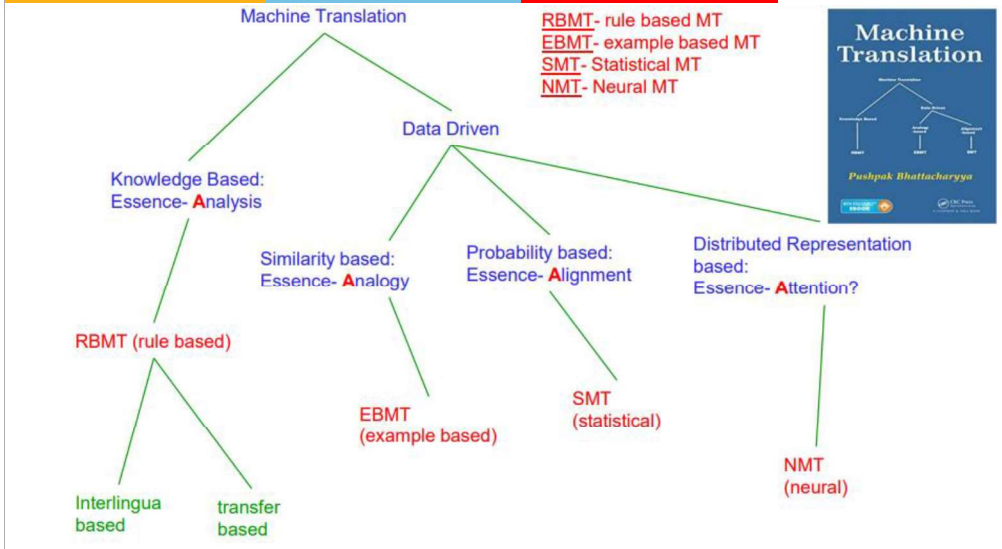
Ref: AI4Bharat Tutorial at ICON 2021, Dec 2021

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 chalaaoing")
- **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.

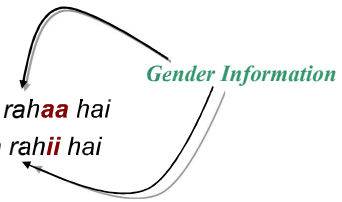
Challenges in Indian Language translation

- **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*

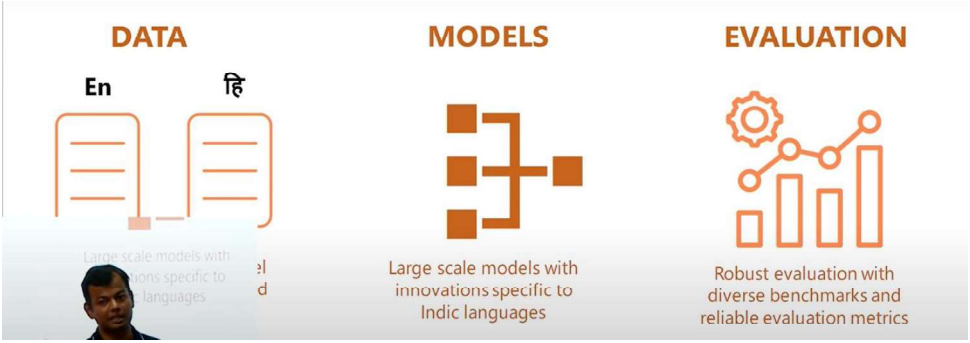
***She** is going to Delhi* ==> ***vaha** dillii jaa rah**ii** hai*

It** broke* ==> ***vaha** TuuTa **??

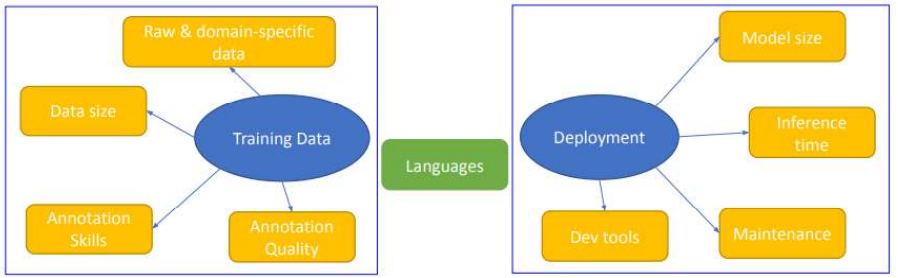
Information does not always map fully from one language into another

Conceptual worlds may be different

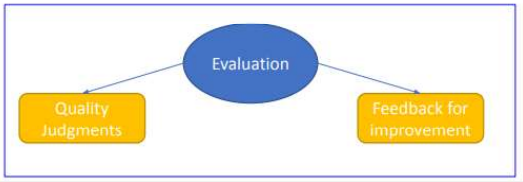
Modern NMT



Scalability Challenges

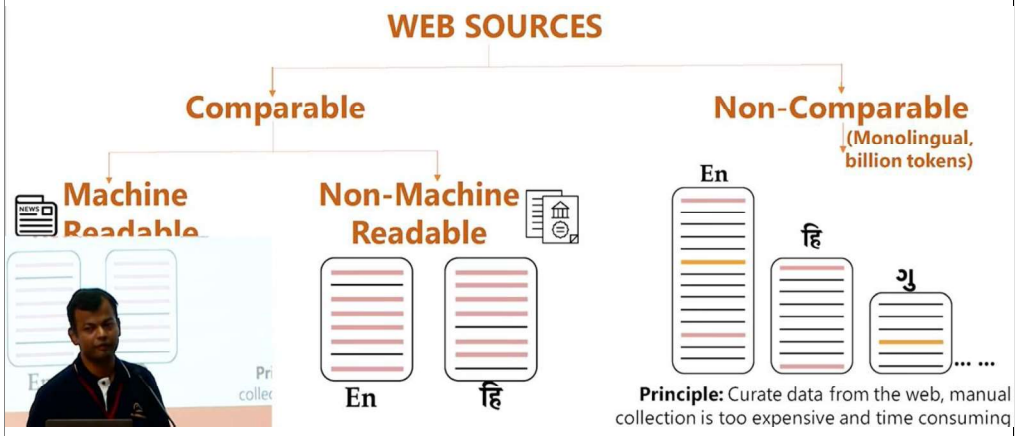


Effort and cost increase as languages increase

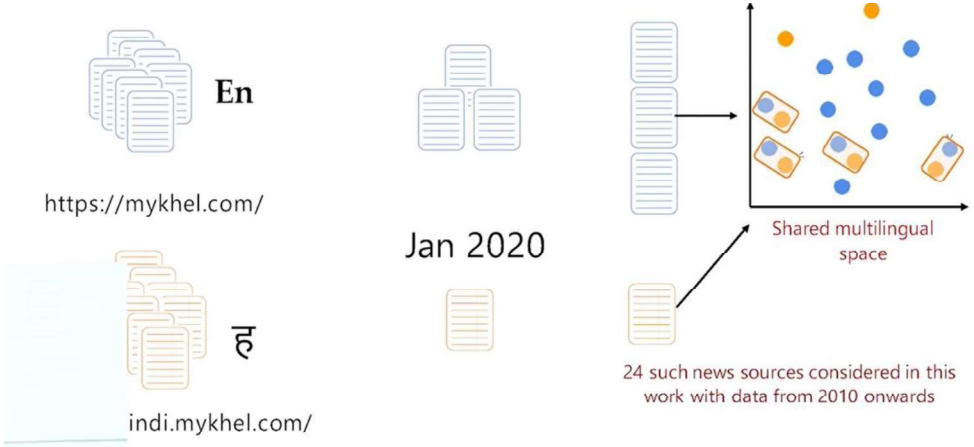


Ref: AI4Bharat Tutorial at ICON 2021, Dec 2021

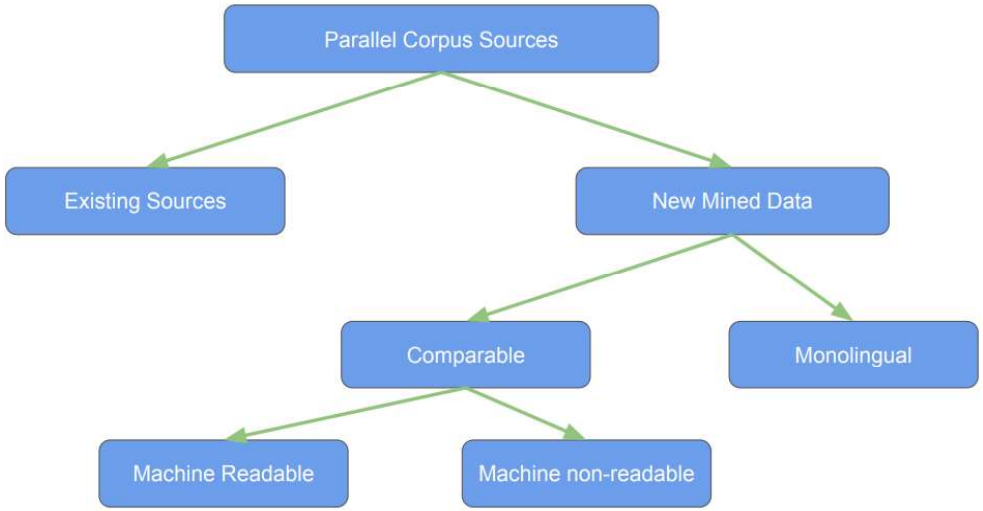
How to solve data problem



How to solve data problem



Parallel Corpora Collection for 11 Indic Languages

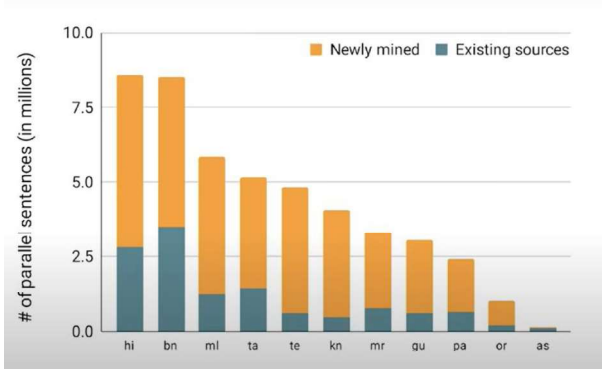


Mining from Machine Readable Sources

- 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

How much data collected

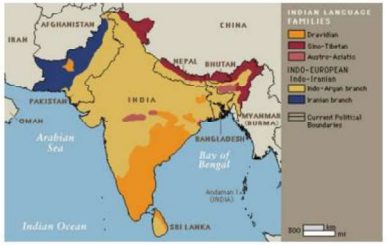


33M parallel sentences mined from web
3 X times than earlier

Neural machine Transliteration

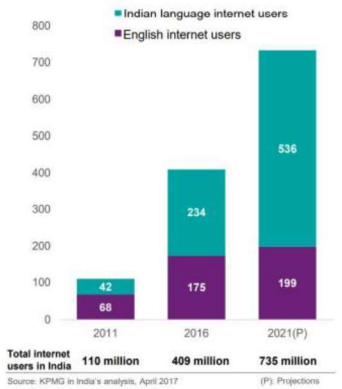
Enable Romanized typing in Indian Languages

Usage of Indian Languages



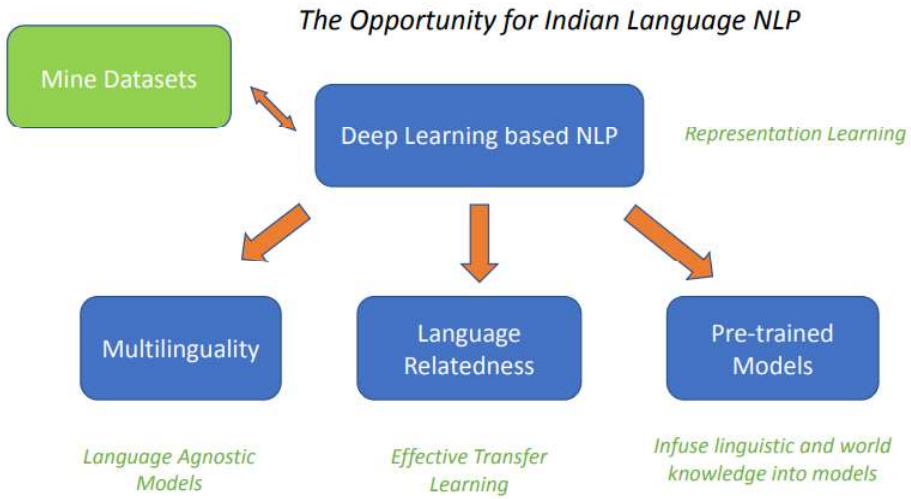
- *4 major language families*
- *22 scheduled languages*
- *125 million English speakers*
- *8 languages in the world's top 20 languages*
- *30 languages with more than 1 million speakers*

Sources: Wikipedia, Census of India 2011



Internet User Base in India (in million)

Source: Indian Languages: Defining India's Internet KPMG-Google Report 2017



ML Approach

Simple Features

Bag-of-words (presence/absence)

Well-made	hit	script	lovely	boring	music
1	1	1	1	0	1

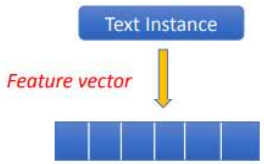
Large and sparse feature vector: size of vocabulary
Each feature is atomic ☐ similarity between features, synonyms not captured

More features

- Bigrams: e.g. *lovely_script*
- Presence in [positive/negative] sentiment word list
- Negation words
- Is the sentence sarcastic (output from saracasm 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)*

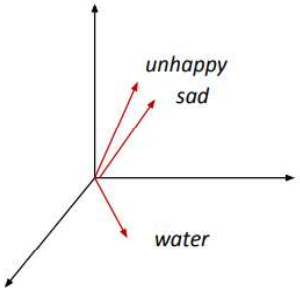
"Words that occur in similar contexts tend to have similar meanings"
- Turney and Pantel (2010)

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

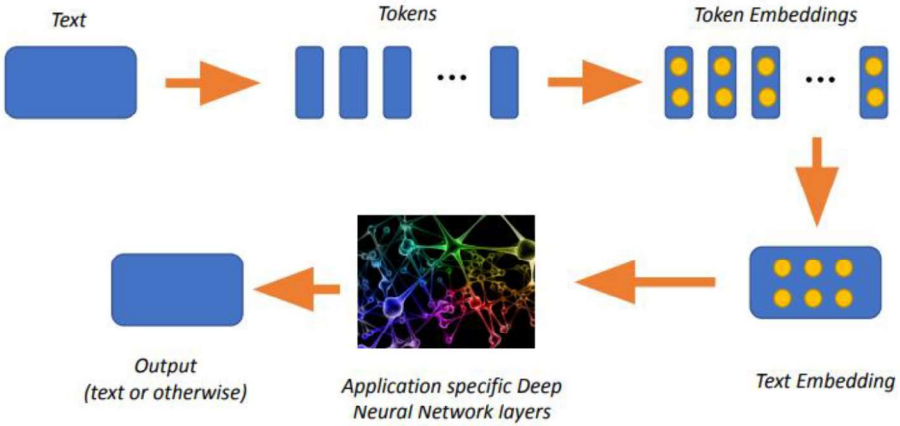


Similarity of words can be defined in terms of vector similarity: Cosine similarity, Euclidean distance, Mahalanobis distance

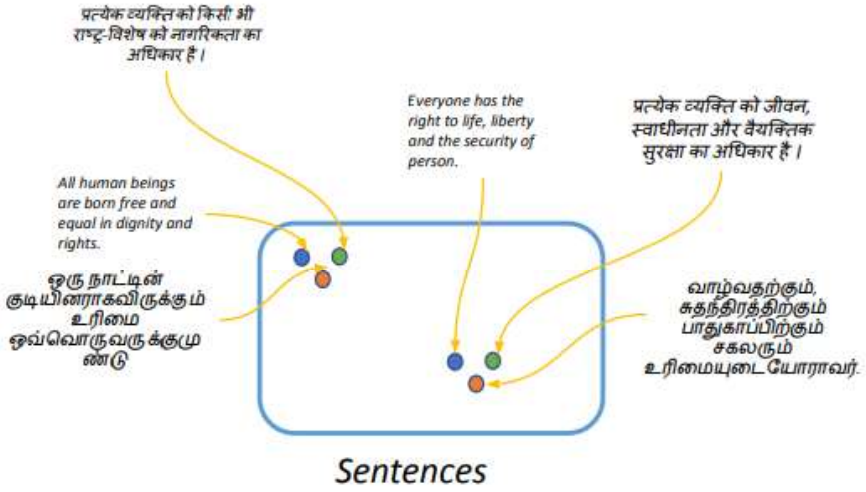
Similarity across languages

Contextual representation of words

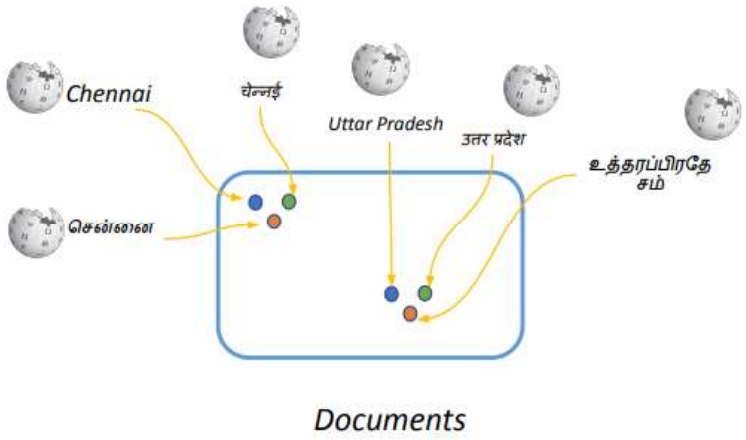
Deep Learning Pipeline



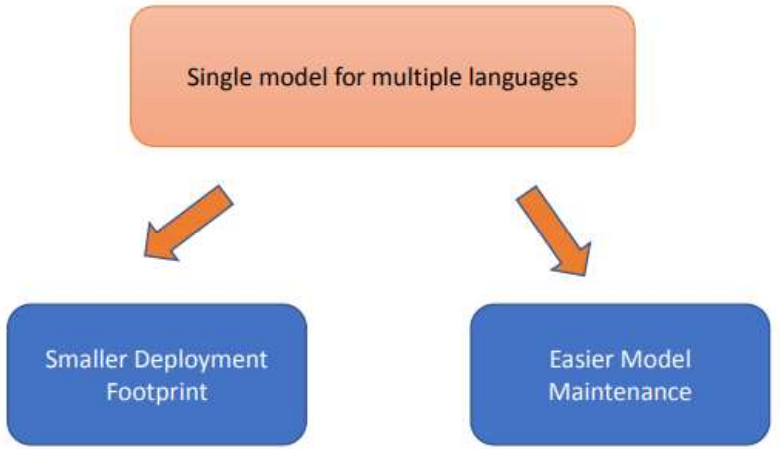
Represent semantically similar language artifacts in the same vector space



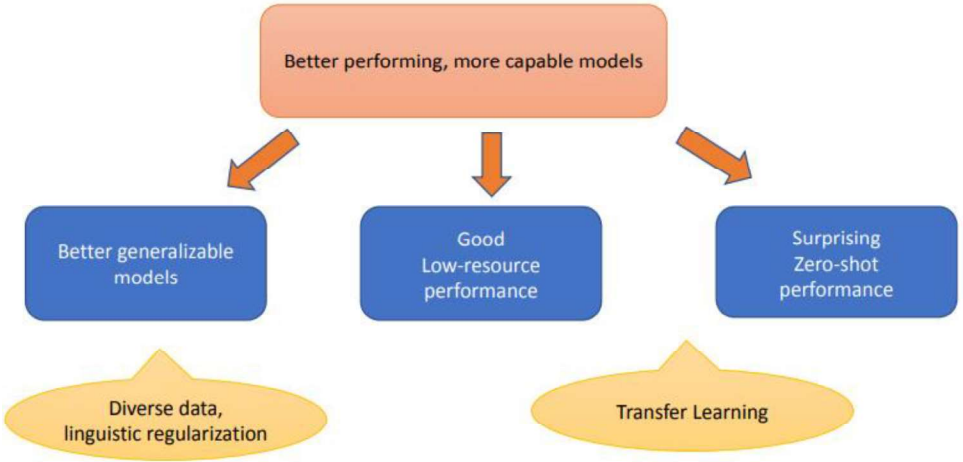
Represent semantically similar language artifacts in the same vector space



Multilinguality



Multilinguality



Encoder Decoder Model

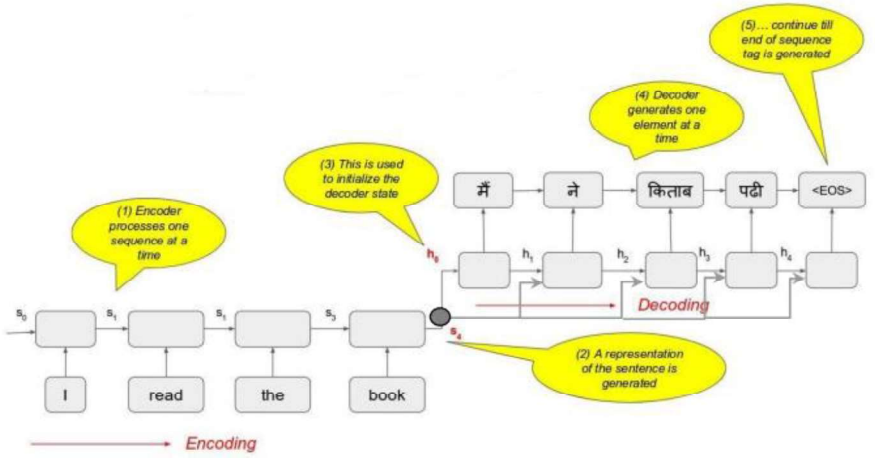
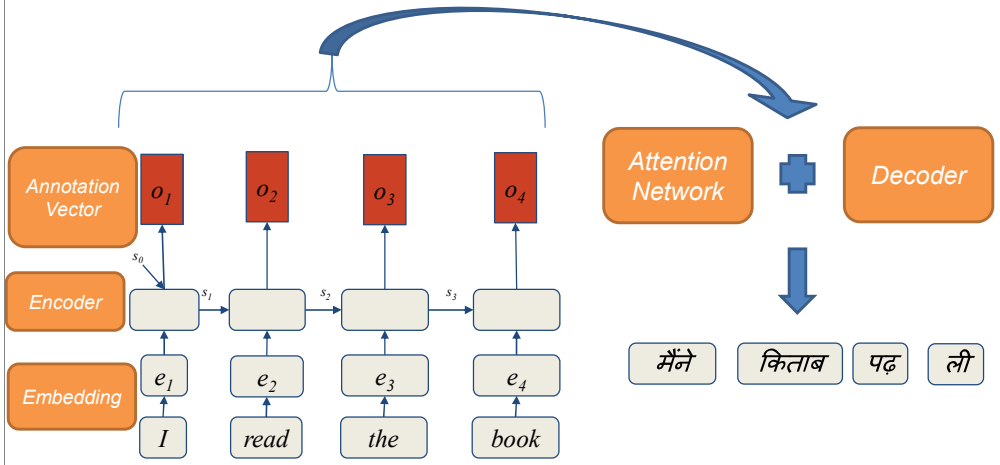


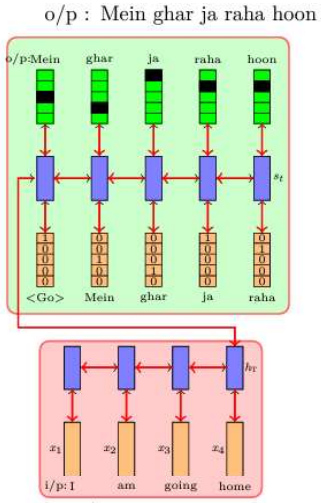
Image source- http://www.iitp.ac.in/~shad.pcs15/data/nmt-rudra.pdf

Embed - Encode - Attend - Decode Paradigm

(Bahdanau et al, 2015)



Encoder Decoder Architecture for Indic MT



- **Task:** Machine translation
- **Data:** $\{x_i = source_i,\ y_i = target_i\}_{i=1}^{N}$
- **Model (Option 1):**
  - **Encoder:**

$$h_t = RNN(h_{t-1}, x_{it})$$

  - **Decoder:**

$$s_0 = h_T \quad (T \text{ is length of input})$$

$$s_t = RNN(s_{t-1}, e(\hat{y}_{t-1}))$$

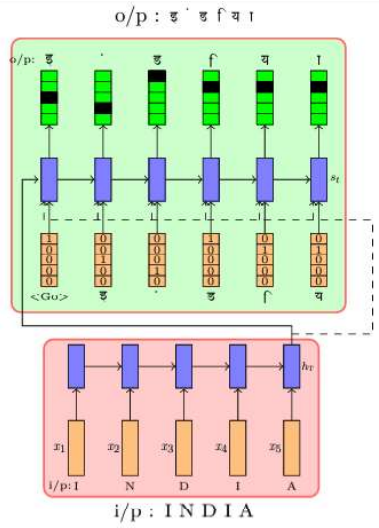
$$P(y_t|y_1^{t-1}, x) = softmax(Vs_t + b)$$

- **Parameters:** $U_{dec},\ V,\ W_{dec},\ U_{enc},\ W_{enc},\ b$
- **Loss:**

$$\mathscr{L}(\theta) = \sum_{t=1}^{T} \mathscr{L}_t(\theta) = -\sum_{t=1}^{T} \log P(y_t = \ell_t|y_1^{t-1}, x)$$

- **Algorithm:** Gradient descent with backpropagation

Encoder Decoder Architecture for Transliteration



Transformer

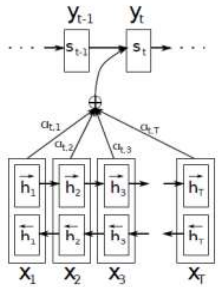
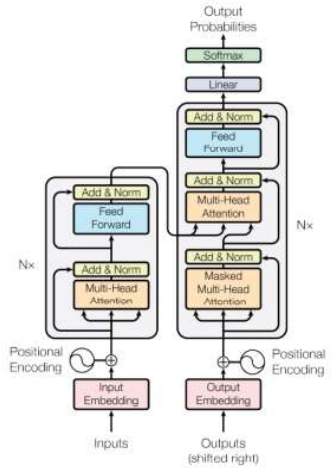


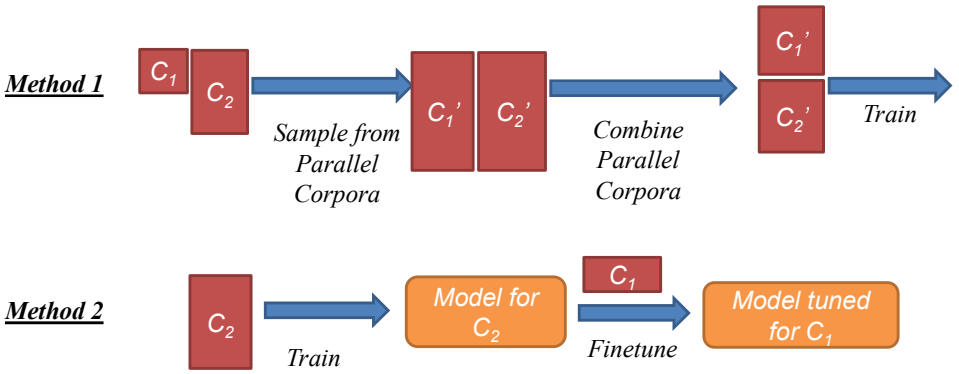
Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

BiLSTM encoder decoder [3]

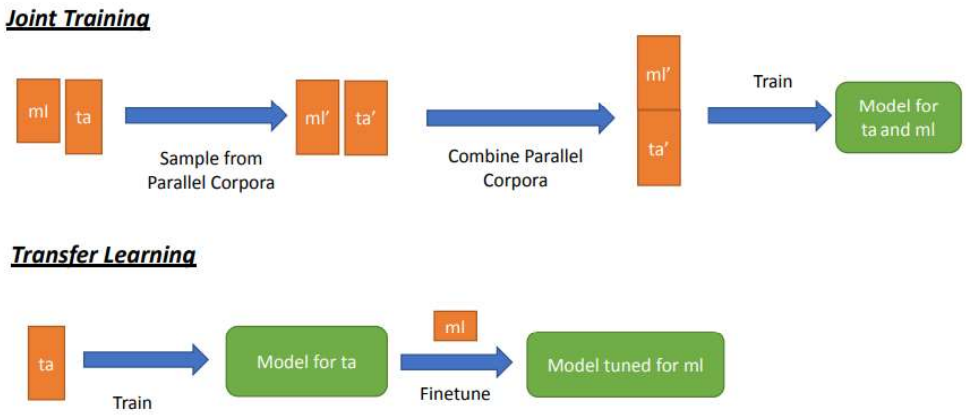


Transformer [8]

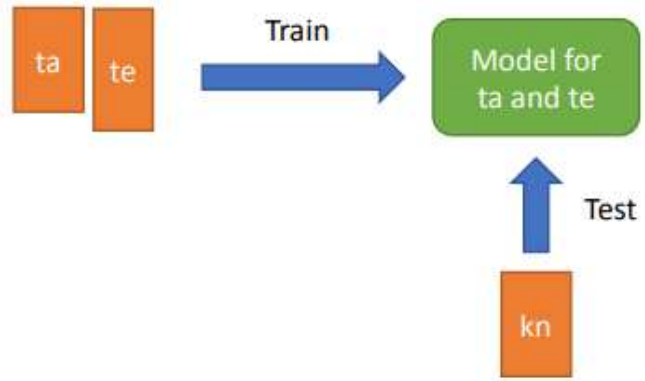
Training Multilingual NMT systems



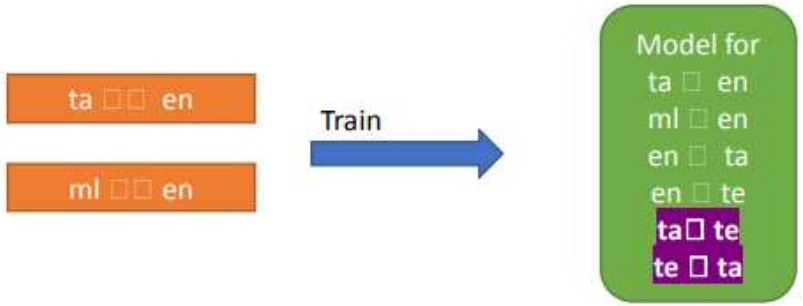
Training multilingual NMT



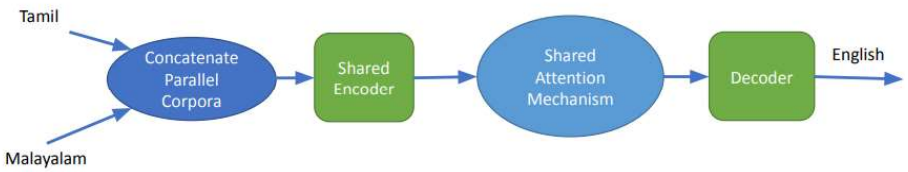
Zero shot translation to English



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

Related Languages

Related by Genealogy → *Related by Contact*

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)

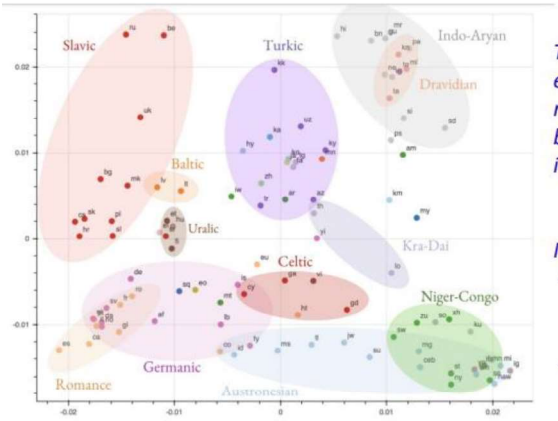
Most translation requirements also involves related languages

Between related languages
Hindi-Malayalam
Marathi-Bengali
Czech-Slovak

Related languages ⇐⇒ Link languages
Kannada,Gujarati ⇒ English
English ⇒ 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 एन्जल्स शहरात कार्यक्रम आयोजित करण्यात आला — Marathi

bhAratAcyA svAtaMtryadinAnimitta amerikeltIla lOs enjalsa shaharAta kAryakrama Ayojita karaNyAta AlA

भारता च्या स्वातंत्र्य दिना निमित्त अमेरिके तील लॉस एन्जल्स शहरा त कार्यक्रम आयोजित करण्यात आला — Marathi segmented

bhAratA cyA svAtaMtrya dinA nimitta amerike tIla lOs enjalsa shaharA ta kAryakrama Ayojita karaNyAta AlA

भारत के स्वतन्त्रता दिवस के अवसर पर अमरीका के लॉस एन्जल्स 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

- 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

$$unicode\_codepoint(char) - Unicode\_range\_start(L_1) + Unicode\_range\_start(L_2)$$

केरला kerala

केरला ಕೆರಳ

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*

- *Loan Words*

borrowed without translation

matsya (sa)	*matsyalu (te)*	*fish*
pazha.m (ta)	*phala (hi)*	*fruit*

- *Named Entities*

do not change across languages

mu.mbal (hi)	mu.mbal (pa)	mu.mbal (pa)
keral (hi)	k.eraLA (ml)	keraL (mr)

- *Fixed Expressions/Idioms*

MWE with non-compositional semantics

dAla me.n kuCha kAlA honA	*(hi)*	*Something fishy*
dALa mA kAlka 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	***Encoder Language Model for NLU***	***Decoder Language Model for NLG***	***Encoder-decoder Language Model for NLU+NLG***
fastText	Google BERT	OpenAI GPT-3	FAIRSEQ BART
	✚ *Multilinguality*		
MUSE	mBERT		mBART

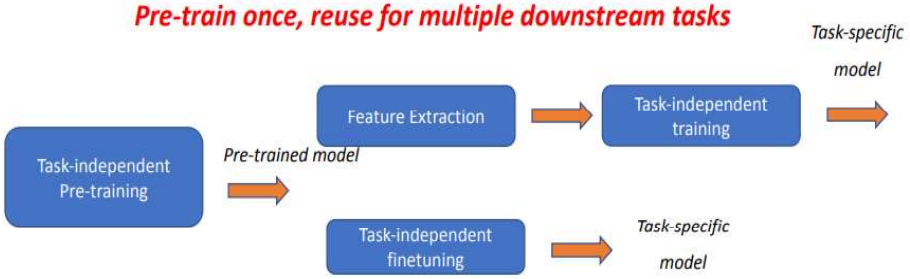
Trained on a large amount of raw text corpora with unsupervised objectives

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

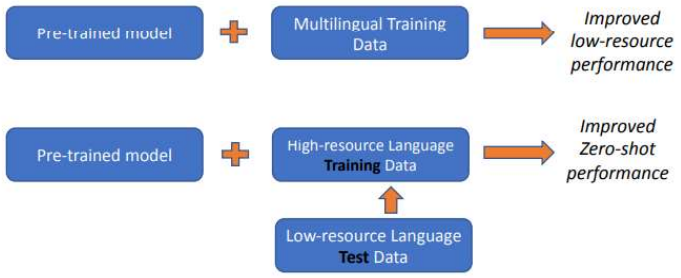
Language modeling & Language generation for tasks like summarization, ASR, question generation

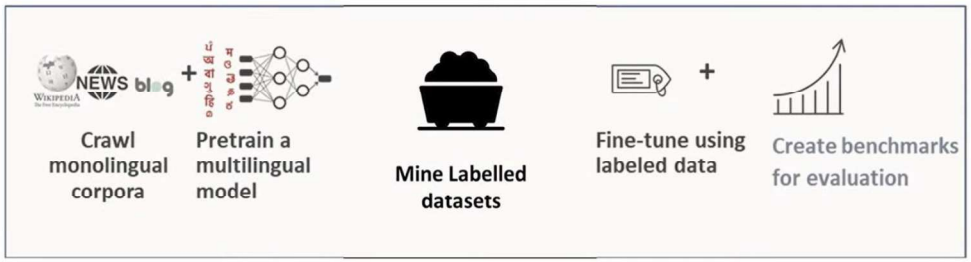
Multilinguality and NN Models

Multi-linguality and Pre-training are complementary

Language-family specific pre-trained model

- *Compact pre-trained models*
- *Utilize language relatedness*
- *Better data representation*



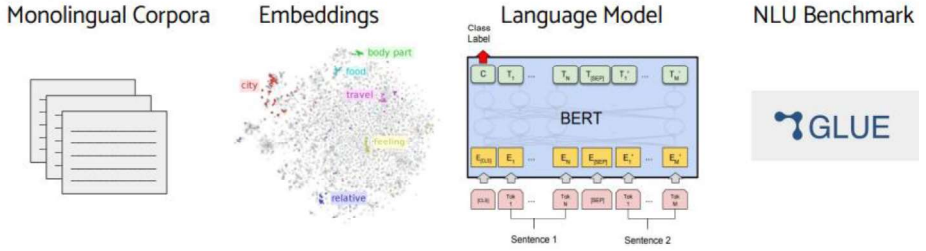


IndicCorp *IndicBERT* *IndicGLUE* *IndicNER*

Naampadam

IndicBART *Indic NLG Benchmark*

IndicNLPSuite



IndicCorp IndicFT IndicBERT IndicGLUE

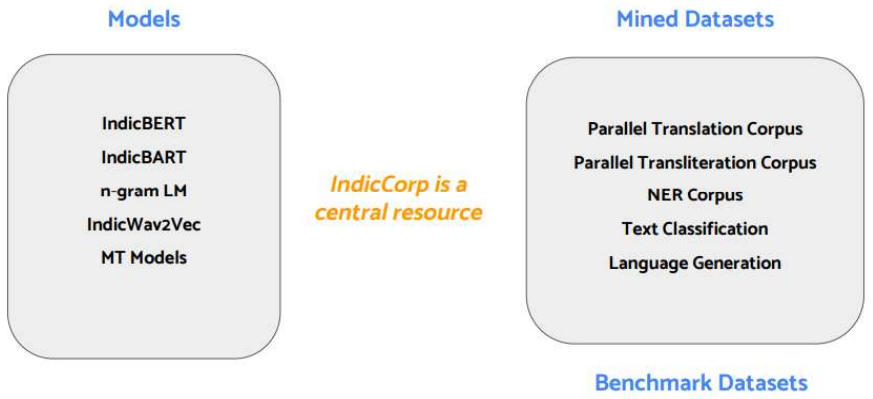
Large scale Monolingual Corpora → **IndicCorp** **450M** Sents.

Evaluation benchmarks → **IndicGLUE** **11** Tasks

Coming soon for IN-22

Multilingual Language Model for IN-22 → **IndicBERT** **18M** Parameter Model

Models and Benchmark datasets



Processing Web content

Processing HTML Pages to Get Sentences

https://github.com/AI4Bharat/webcorpus

HTML Page → (1.) → Extracted Article → (2., 3.) → Extracted Sentences

```
<html>
<head>...</head>
<body>...</body>
</html>
```

RCB batting coach Sanjay Bangar, ahead of their contest against a struggling Punjab Kings...

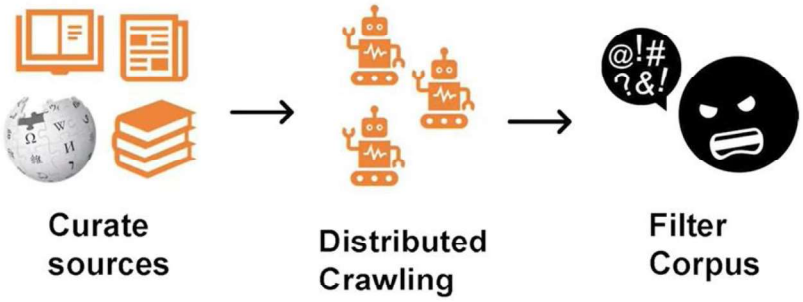
RCB appear to be a well-oiled machine in this season of IPL 2021

There are multiple players who on their day can win the match on their own

Boilerpipe
Custom Extractors

IndicNLP library
Filters (language, length,script)

Monolingual Corpora collection



https://github.com/AI4Bharat/webcorpus

Multilingual Word Embedding

மரத்தாலான (wooden)

மரத்தால் (tree) + ஆன (making)

Complex tense, verb embedded into a single word

Indic FastText

IndicGlue - Bencmarks

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

Unsupervised Mining

Task 1

Task 2

Task 3

IndicGlue Tasks

6 Tasks
4 Types

Semantic
News Articles Headline Prediction — News Crawls
Wikipedia Section Title Prediction — Wikipedia
Article Genre Classification — 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

Task: Predict the correct headline

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

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

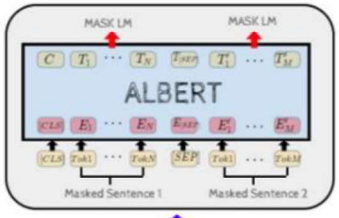
Created From: News Crawl **Task**: Predict the genre of news article

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

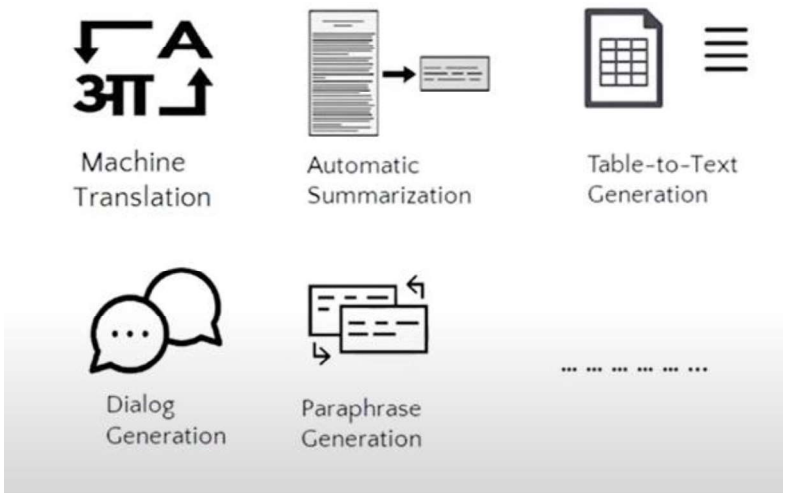
IndicBERT



- 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

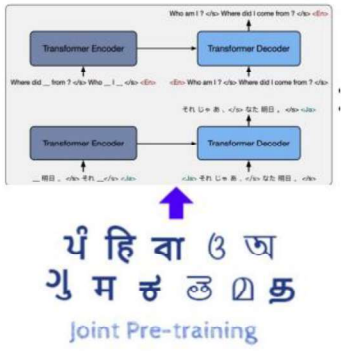
ai4bharat/indic-bert · Hugging Face

Natural Language Generation



Missing in Indic Languages

IndicBART



- 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

நான் ஒரு பையன் → नान् ओरु पैयन्

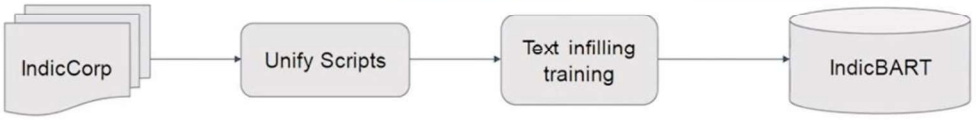
ഞാൻ ഒരു പയ്യൻ → न्यान् ओरु पय्यन्

IndicBART Training

- Train models to do: (text infilling)

मुझे [MASK] पसंद हैं। → मुझे भाषांतर पसंद हैं।

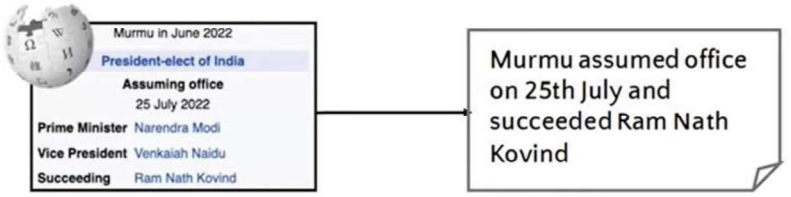
I [MASK]. → I love translation.



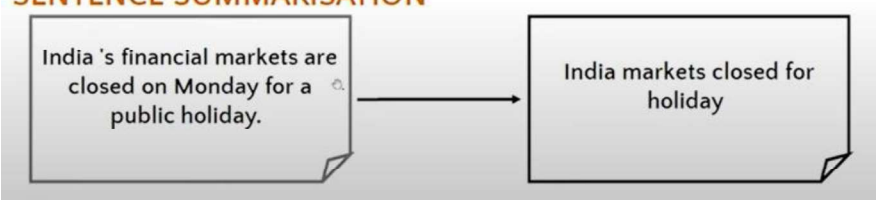
- IndicBART learns to infer a variation of input.
  - Learns generic NLG →Reduces need for task data (fine-tuning)
  - Variations: IndicALBART (compact)

Methods for creating training data

BIOGRAPHY GENERATION



SENTENCE SUMMARISATION

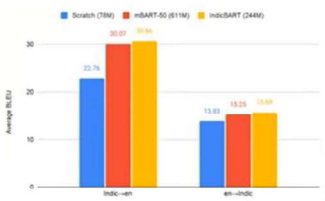


Methods for creating training data

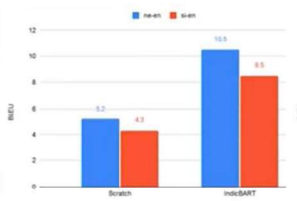
PARAPHRASE GENERATION

QUESTION GENERATION

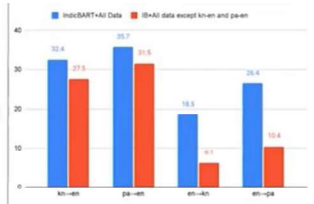
Machine Translation Bleu Score



- Large impact of pre-training
  - *Indic→En: 22.76→30.66*
  - *En→Indic: 13.83→15.69*
- Indic→En gains more than En→Indic

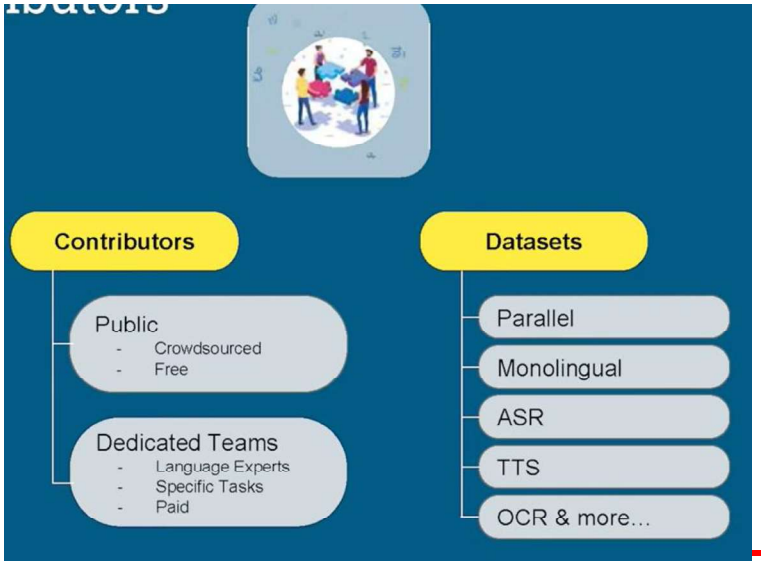


- IndicBART helps Nepali and Sinhala translation
- Both were unseen by IndicBART

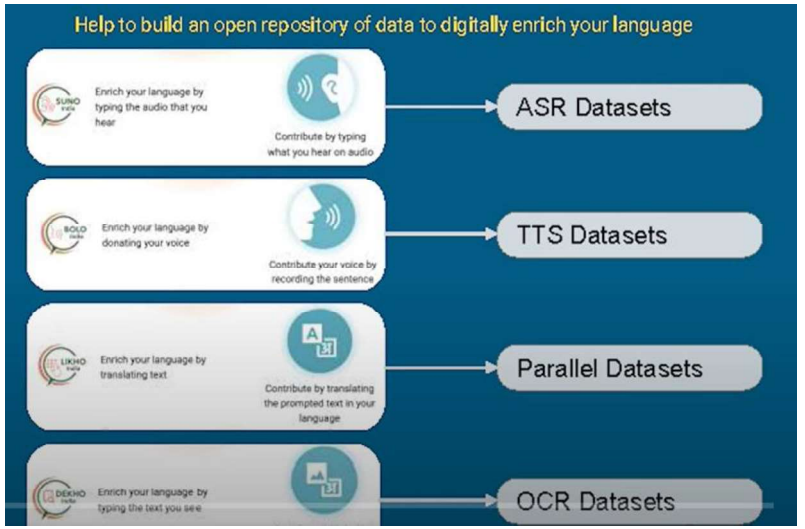


- IndicBART helps unseen language translation
- Punjabi and Kannada data not used
  - Can still translate

Datasets



Bhasha daan



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
- Multiligual Neural MT
- https://www.youtube.com/watch?v=BdZeN-6TYzs
- https://www.youtube.com/@ai4bharat/featured

Code References

https://bhashini.gov.in/ulca/search-model/65[illegible]d0f00d64169e2f8f3f8/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, Ml, Mr, Or, Pa, Ta, Te
    - Automated Creation via entity projection
  - Human annotated test-set for 8 Indic languages
    - Bn, Hi, Kn, Ml, Mr (large)
    - Ta, Te, Gu (small)
- Multilingual IndicNER model
  - 11 Indic languages (As, Bn, Gu, Hi, Kn, Ml, 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

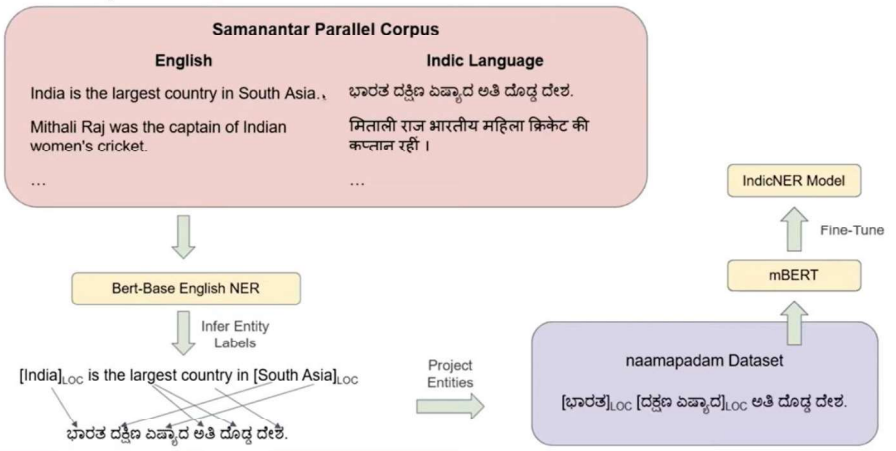
For example,

[Nilekani Center]$_{LOCATION}$ at [AI4Bharat]$_{ORGANIZATION}$ will be launched on [28th July]$_{DATE}$ at [IIT Madras]$_{ORGANIZATION}$

Challenges in Indic languages:

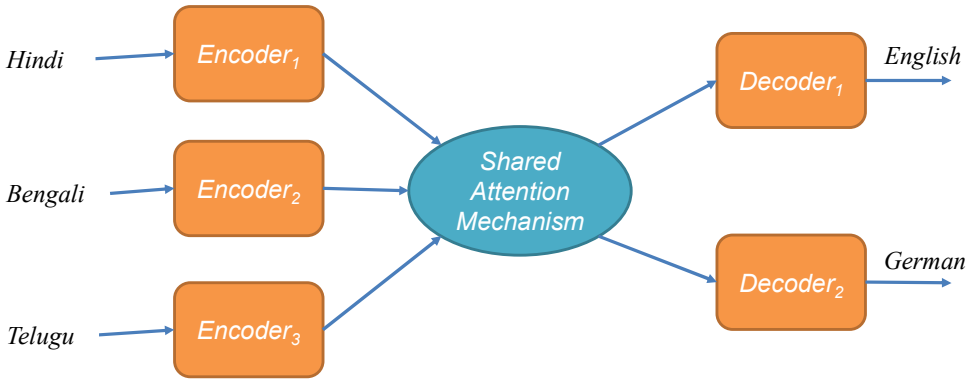
- Lack of capitalization feature
- Ambiguity between Proper nouns and common nouns
- Morphological variations
- Small labelled data

Indic NER Model



Multilingual Neural Translation

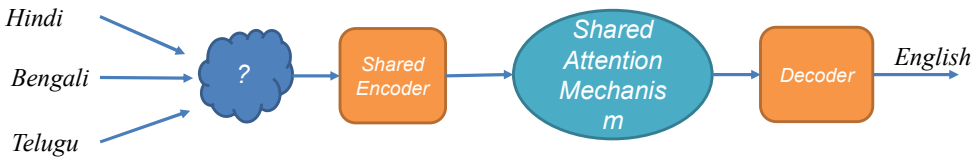
(Firat et al., 2016; Johnson et al., 2017)



Translate unseen language pairs ➔ Zeroshot Translation

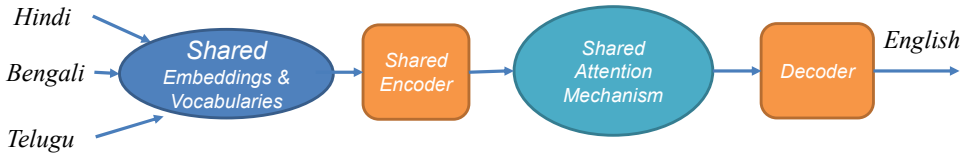
Shared Encoder

(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017)



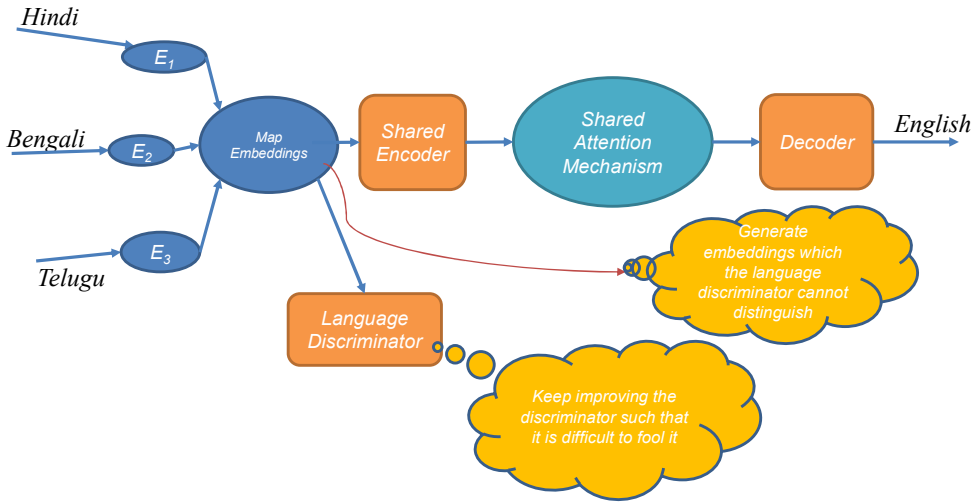
Shared Encoder

(Zoph et al., 2016; Nguyen et al., 2017; Lee et al., 2017)

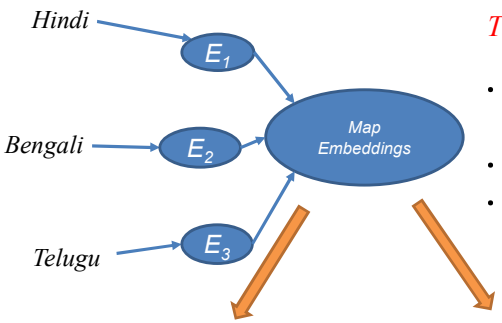


Shared Encoder with Adversarial Training

(Joty et al., 2017)



Learning Multilingual mappings/embeddings



The key to multilingual learning

- *Text Classification (sentiment analysis, Question matching)*
- *Sequence Tagging (POS, NER, etc.)*
- *Sequence to Sequence Learning (Machine Translation, Transliteration, etc)*

Offline Mapping of embeddings

$e_2 = A_{21}e_1$

Joint training for multilingual embeddings

$A_1e_1 = A_2e_2$

Needs parallel corpora or bilingual dictionaries

Tutorial on Multilingual Multimodal Language Processing Using Neural Networks at NAACL 2016, Mitesh Khapra & Sarath Chandar

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

BITS Pilani
Pilani Campus

Session 12: Information Extraction-Named Entity Recognition
Date – 3rd 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

- Information Extraction
- Typical IE Pipeline
- Named Entity Recognition
- Challenges in NER
- NER Approaches
- MEMM
- CRF
- Neural NER
- Evaluation of NER

Information Extraction (IE)

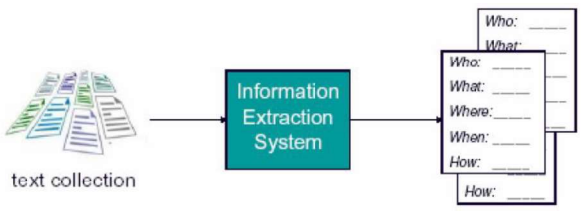
- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
  - Newspaper articles
  - Web pages
  - Scientific articles
  - 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…

IE

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…

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…

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…

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.

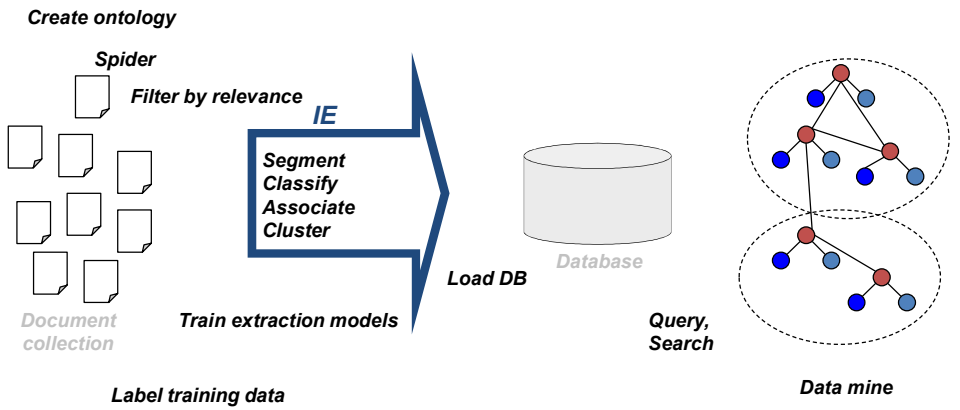
"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Adapted from slide by William Cohen

IE in Context



Adapted from slide by William Cohen

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.

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

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.			

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 *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	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 *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	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

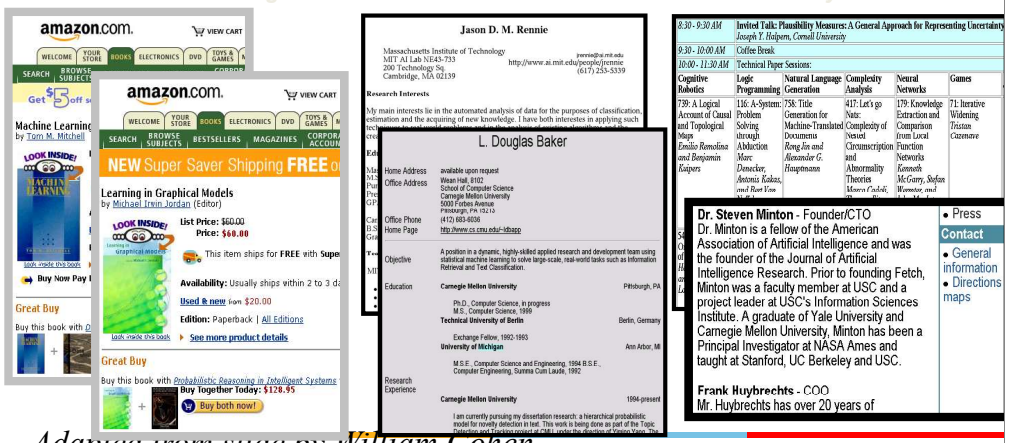
Adapted from slide by William Cohen

Landscape of IE Tasks:
Intended Breadth of Coverage

Web site specific
Formatting
Amazon.com Book Pages

Genre specific
Layout
Resumes

Wide, non-specific
Language
University Names



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

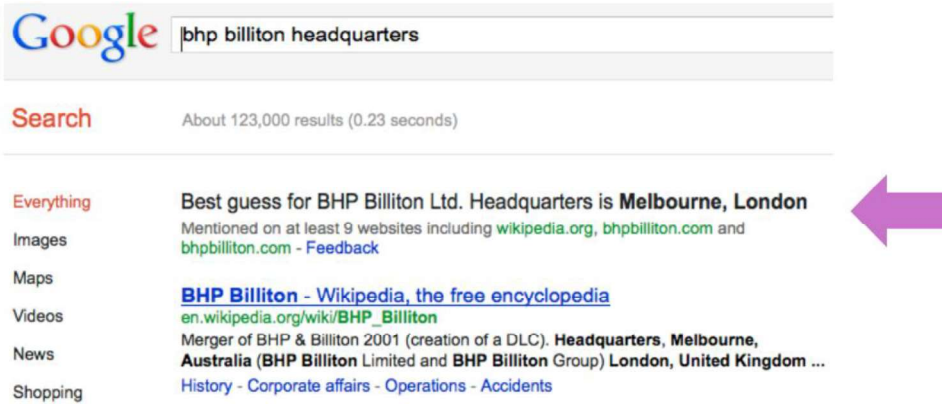
- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- E.g.,
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are 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



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
Organi-
zation

Evolution of NER

Traditional	Statistical	Neural
• Rules • Regular Expressions • Gazetteers	• Word-based models – PMI, log-likelihood. • Sequence models – Conditional Random Fields	• 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
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

} Standard evaluation is per entity, *not* 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

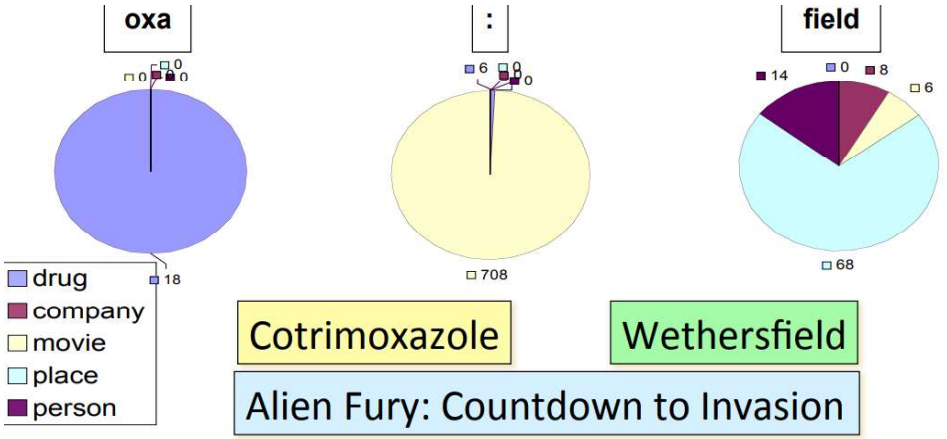
Encoding classes for sequence labeling

	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

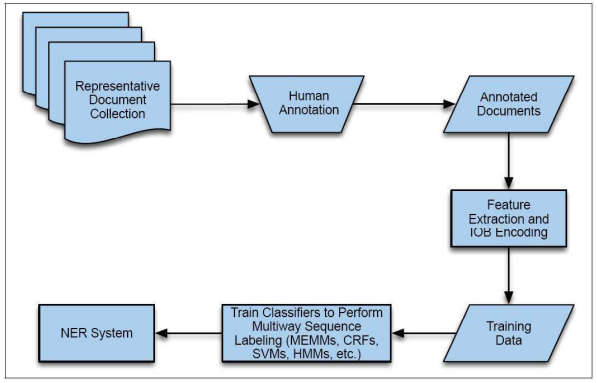
Named Entity Recognition

[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	*T98G, HeLa cell, Chinese hamster ovary cells, CHO cells*
Cell types	*primary T lymphocytes, natural killer cells, NK cells*
Chemicals	*citric acid, 1,2-diiodopentane, C*
Drugs	*cyclosporin A, CDDP*
Genes/proteins	*white, HSP60, protein kinase C, L23A*
Malignancies	*carcinoma, breast neoplasms*
Medical/clinical concepts	*amyotrophic lateral sclerosis*
Mouse strains	*I.AFT, AKR*
Mutations	*C10T, Ala64 → Gly*
Populations	*judo group*

Source: Jurafsky & Martin "Speech and Language Processing"

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)

Source: Jurafsky & Martin "Speech and Language Processing"

Sequence Problems

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

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

POS tagging

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

Named entity recognition

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

Word segmentation

Text segmentation

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	%

(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

- 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

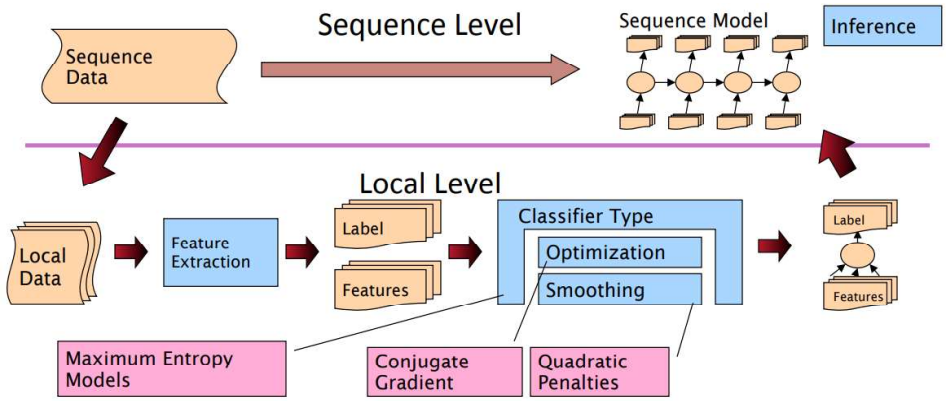
-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

(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: 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

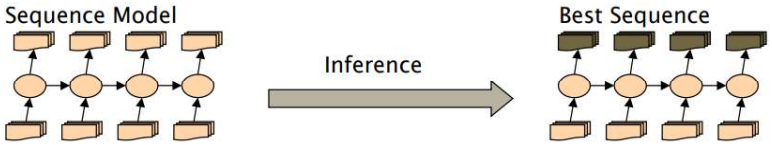
Advantages:

- 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.

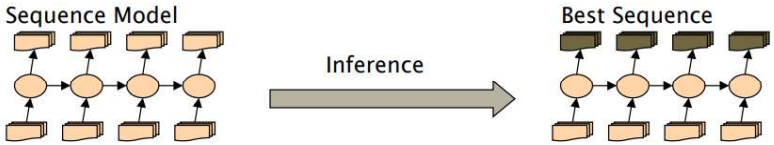
Advantages:

- 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.

$$P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

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 Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
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'Xxxxxxxx")
- Short Word Shape: remove adjacent letters that are identical in word shape "L'Occitane → "X'Xxxxxxxx" → "X'Xx")

Input Format – BIO Tagging

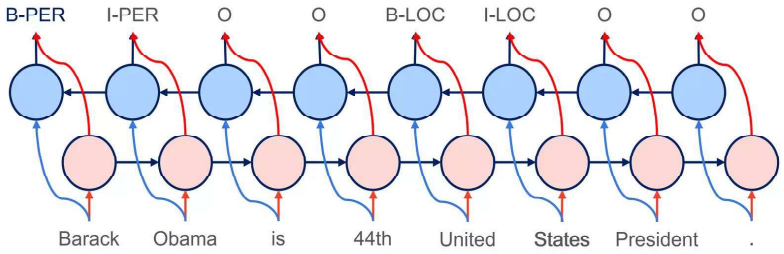
Barack Obama (PER) is 44th United States (LOC) President .

- **BIO** – **B**egin **I**n **O**ut.
  - 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

CoNLL

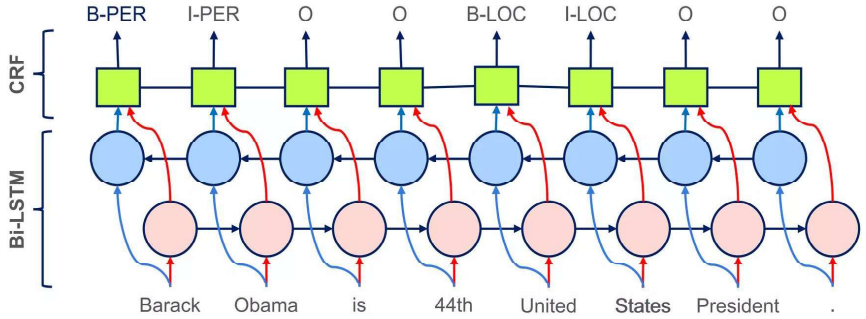
```
Barack      B-PER
Obama       I-PER
is          O
44th        O
United      B-LOC
States      I-LOC
President   O
.           O
```

Neural Model - BiLSTM



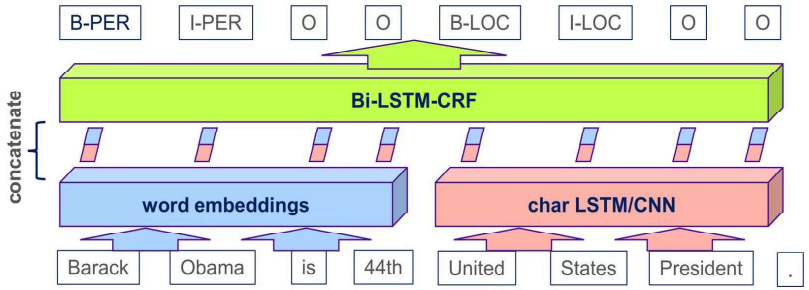
- Input is sequence of tokens, output is sequence of BIO tags.
- Weights trained end-to-end, no feature engineering needed.
- Bidirectional LSTM gets signal from neighboring words on both sides.

Neural Model – BiLSTM-CRF



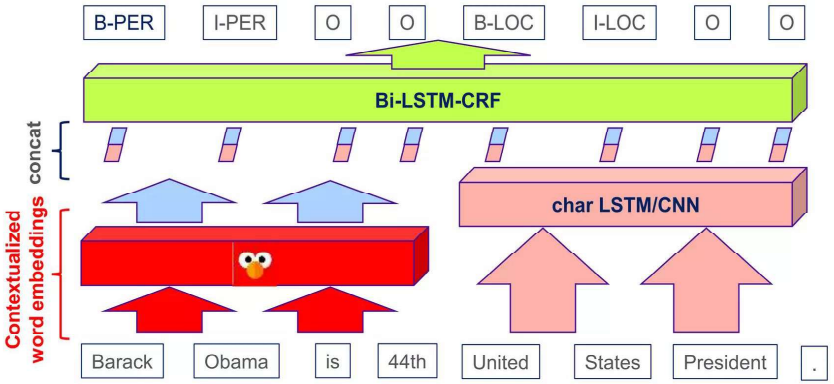
- Same as previous model, with additional CRF layer.
- No feature engineering for CRF, unlike CRF only NER model.
- Pre-trained embeddings observed to improve performance.

Neural Model – adding char embeddings

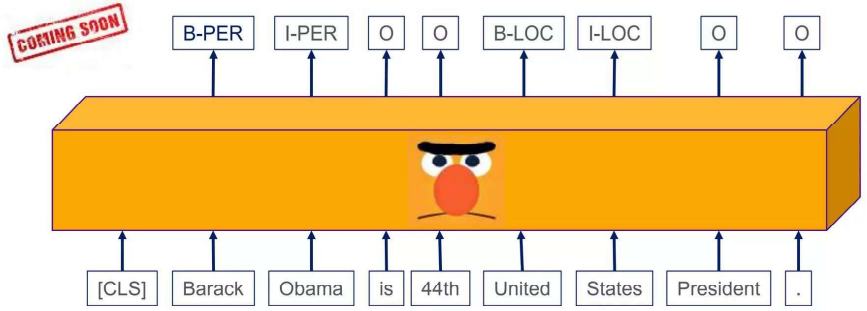


- Concatenate char embedding + word embedding and feed to Bi-LSTM-CRF.
- All weights learned end-to-end.
- Handles rare / unknown words; Exploits signal in prefix/suffix.

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.

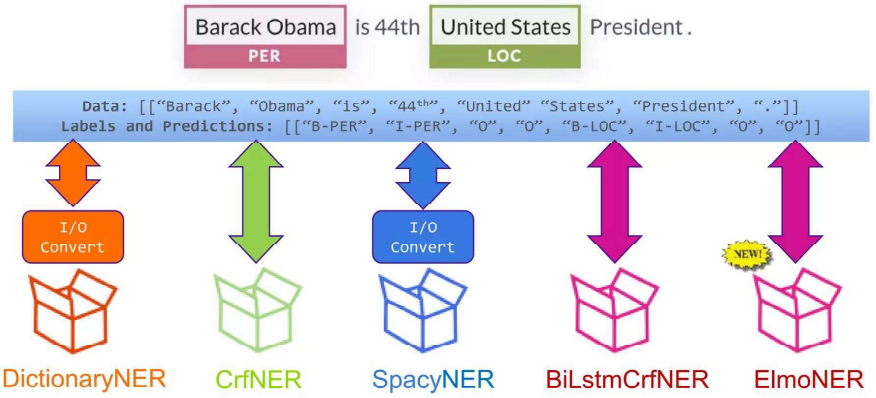
Benefits of Unification

- 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



Ensemble NER

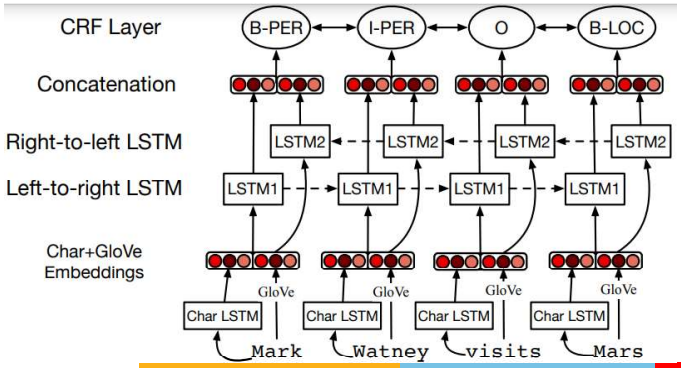
```
# create and test an ensemble
dict_model = DictionaryNER()
dict_model.load("models/dict_model")
crf_model = CrfNER()
crf_model.load("models/crf_model")
spacy_model = SpacyNER()
spacy_model.load("models/spacy_model")
bilstm_model = BiLstmCrfNER()
bilstm_model.load("models/bilstm_model")
model = EnsembleNER()
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

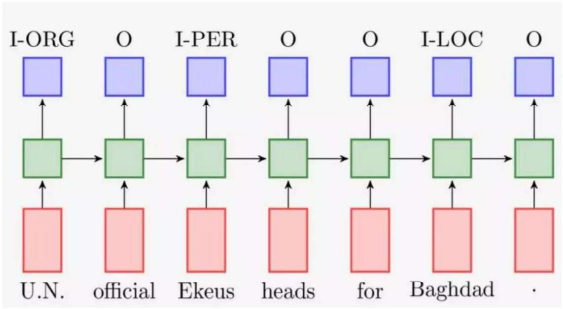
Sequence RNN (e.g. biLSTM or Transformer) with a CRF output layer.

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

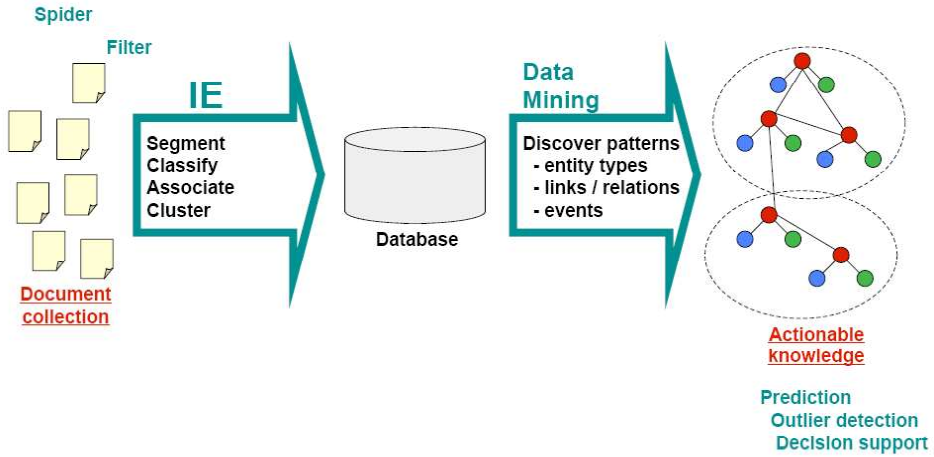
Precision, Recall, F1 for NER

- 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 funnily 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



Source: Andrew McCallum, UMass Amherst

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://slideplayer.com/slide/4235847/
Named entity tagging
https://www.youtube.com/watch?v=7CRyqwCZFY0
NER
https://www.bilibili.com/video/BV1CE41197rQ/?p=36

Natural Language Processing Applications

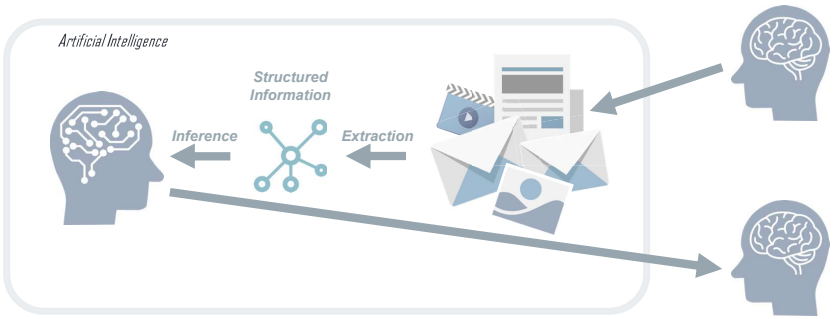
BITS Pilani
Pilani Campus

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

Session 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.

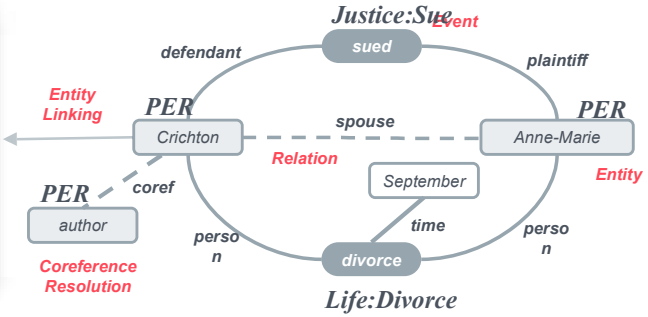
A Quick Overview of Information Extraction



- 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.

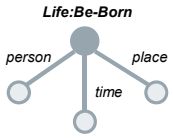
A Quick Overview of Information Extraction

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

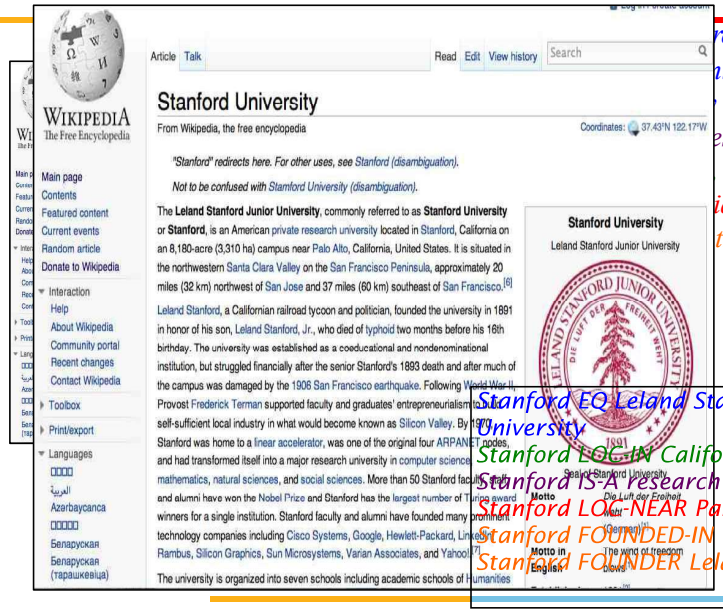
- Company report: "International Business Machines Corporation (IBM or the company) was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)…"
- Extracted Complex Relation:

  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



Why Relation Extraction?

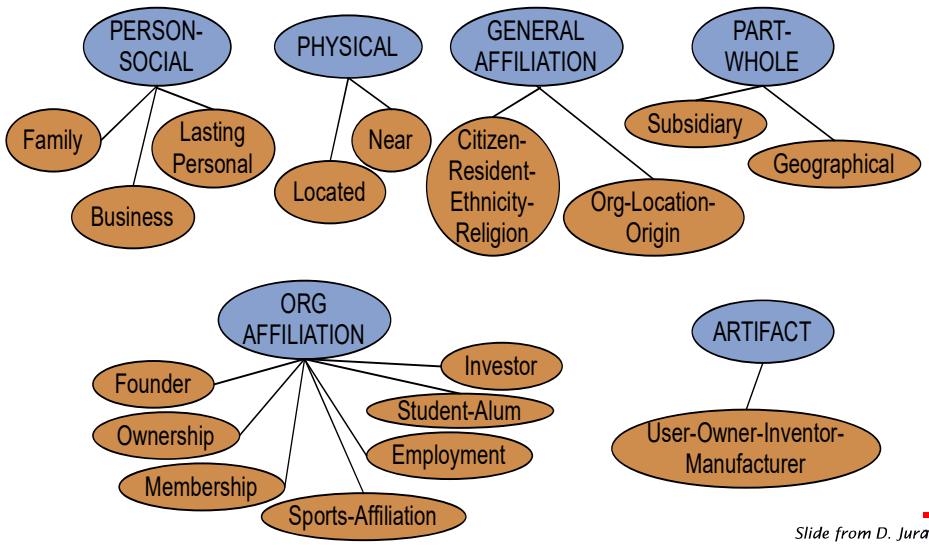
- Create new structured knowledge bases, useful for any app
- Augment current knowledge bases
- Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
- Support ques on answering

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”



Slide from D. Jurafsky

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

Extracting UMLS relations from a sentence

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

↓

Echocardiography, Doppler DIAGNOSES Acquired stenosis

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"}}<br /> ([[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-
05|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 = 2007-04-26}}</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 Frei
…

TEXT PUBLIC

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	→	48-year-old Elon Musk is the founder of SpaceX and a co-founder of Tesla Motors.
X, who founded Y	→	Elon Musk, who founded SpaceX in 2002, has said the company is focused on ...
Y was founded by X	→	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

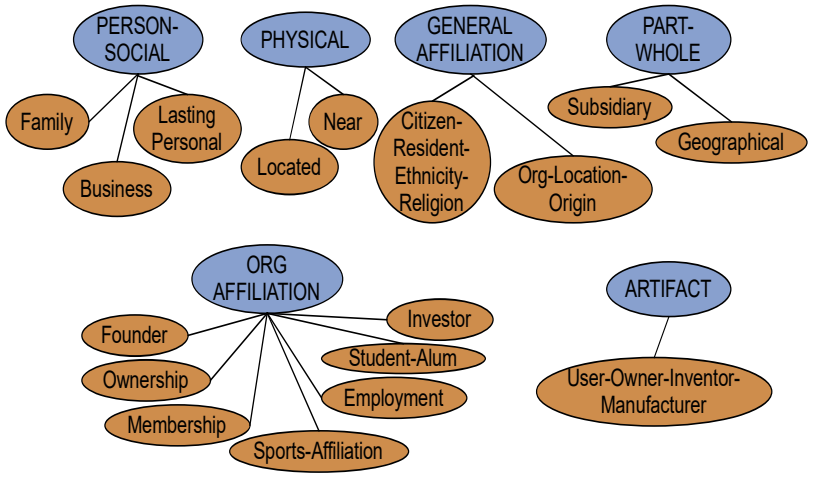
How to do classification in supervised relation extraction

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)

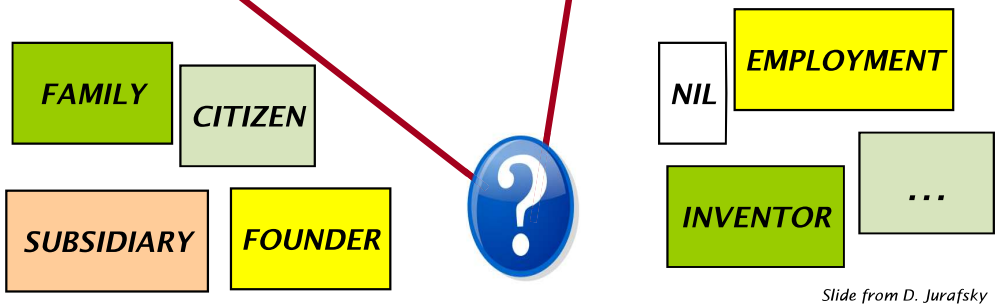
17 sub-relations of 6 relations from 2008 "Relation Extraction Task"



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	*airlines*
$Entity_2$ type	PERS
$Entity_2$ head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before $Entity_1$	NONE
Word(s) after $Entity_2$	*said*
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	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Slide from D. Jurafsky

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

Slide modified from D. Jurafsky

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*

E. Agichtein and L. Gravano 2000. Snowball: Extracting Relations from Large Plain-Text Collections. ICDL

Distant Supervision

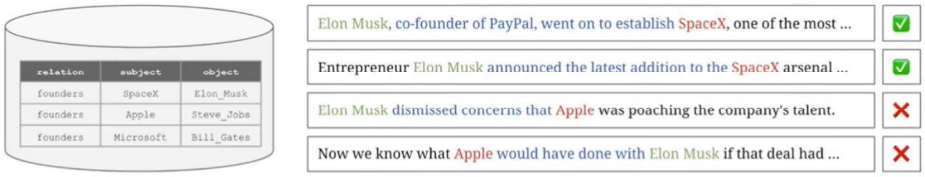
- Combine bootstrapping with supervised learning
  - Instead of 5 seeds,
    - Use a large database to get huge # of seed examples
  - Create lots of features from all these examples
  - Combine in a supervised classifier

Snow, Jurafsky, Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. NIPS 17
Fei Wu and Daniel S. Weld. 2007. Autonomously Semantifying Wikipeida. 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



Hooray! Massive quantities of training data, practically free!

Qualm: are those assumptions reliable?

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

$P(\text{born-in} \mid f_1, f_2, f_3, \ldots, f_{70000})$

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

(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/F_1 for each relation

$$P = \frac{\#\text{ of correctly extracted relations}}{\text{Total \# of extracted relations}}$$

$$R = \frac{\#\text{ of correctly extracted relations}}{\text{Total \# of gold relations}}$$

$$F_1 = \frac{2PR}{P+R}$$

53

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
Boot-strapping	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.

5
6

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.

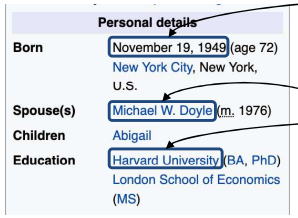
57

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



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

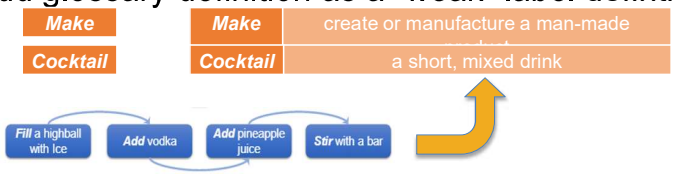
Hoffmann et al. Learning 5000 Relational Extractors. ACL 2010

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 defintion



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

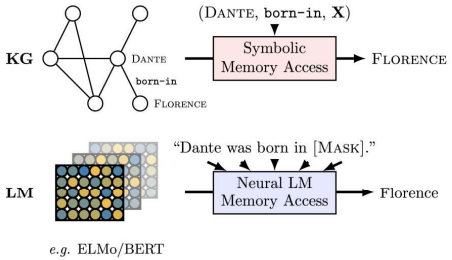
- Pre-trained language models can also be used as distant supervision
  - It did not use additional annotations
  - It is not task-specific
  - It contains inductive biases (weak signals)

Paris is a [MASK].

- PLMs are applied for IE in many creative ways
  - Contextual embeddings to replace word embeddings
  - Direct probing
  - Direct probing + task-specific finetuning

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
TREx	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 Tougui 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

Petroni et al. Language Models as Knowledge Bases?. EMNLP 2019

Weak Supervision

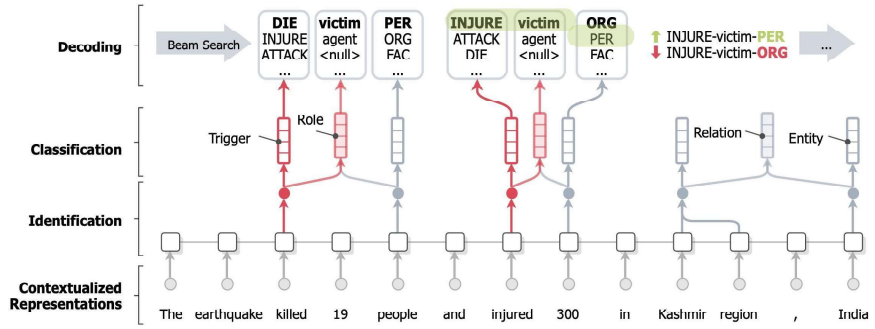
- Mark joined **Amazon** a month ago.
  - What is the entity type?
- Weak Supervision:
  - From knowledge bases
    - Amazon.com, Inc is an American multinational technology company.
  - From weak but richer label representations
    - Word-embedding(company) is close to Word-embedding(Amazon)
  - From pre-trained LMs
    - Amazon is a [MASK] <- [MASK] = company.
  - From linguistic patterns
    - PER *join* company

Weak Supervision – Linguistic Patterns

- 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

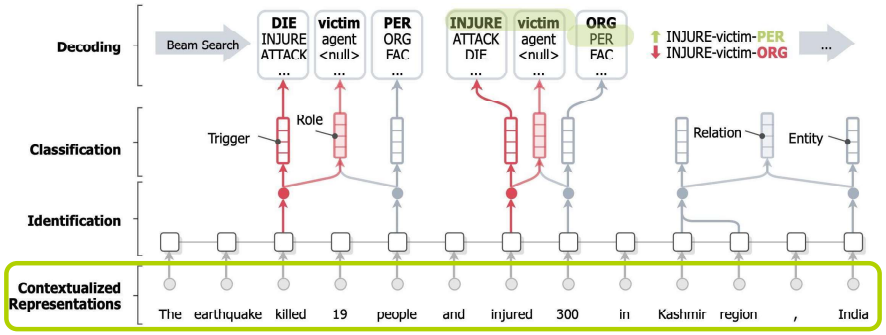
Zhou et al. Temporal Common Sense Acquisition with Minimal Supervision. ACL 2020

OneIE: An End-to-end Neural Model for IE



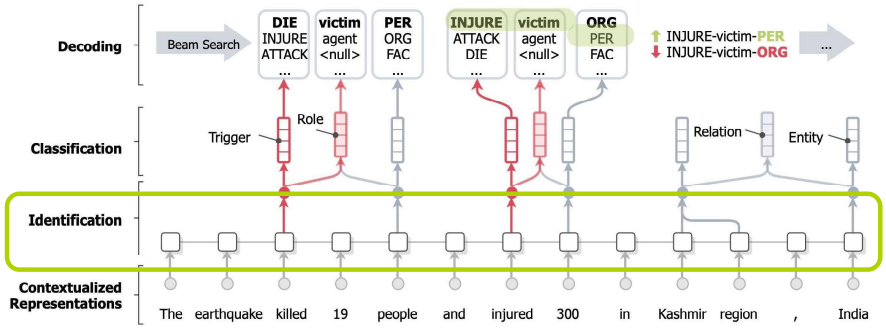
- Our OneIE framework extracts the information graph from a given sentence in four steps: encoding, identification, classification, and decoding

OneIE: An End-to-end Neural Model for IE



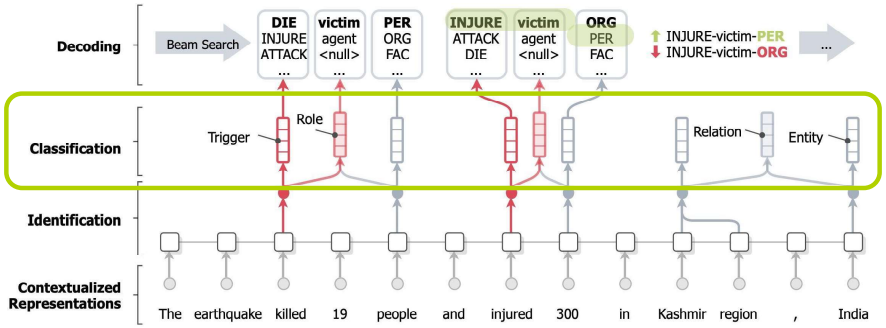
- **Encoding**: We use a BERT encoder to obtain the contextualized embedding of each token

OneIE: An End-to-end Neural Model for IE



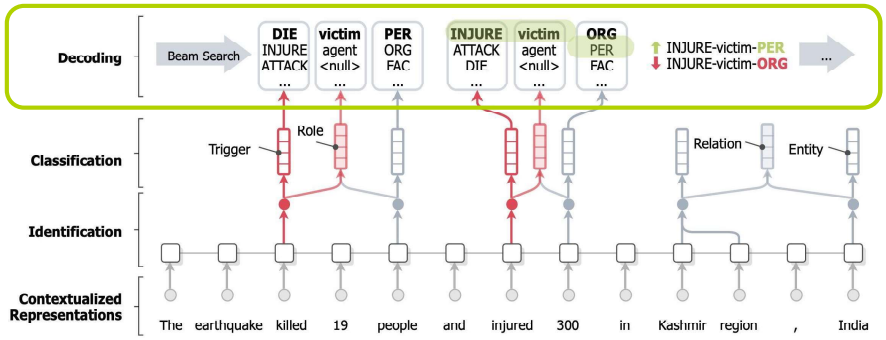
- **Identification**: We use CRF taggers to identify entity mentions and event triggers
- We define the identification loss as $\mathcal{L}^{\mathrm{I}} = -\log p(\boldsymbol{z}|\boldsymbol{X})$

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 $\mathcal{L}^{\mathrm{t}} = -\frac{1}{N^t}\sum_{i=1}^{N^t} \boldsymbol{y}_i^t \log \hat{\boldsymbol{y}}_i^t$

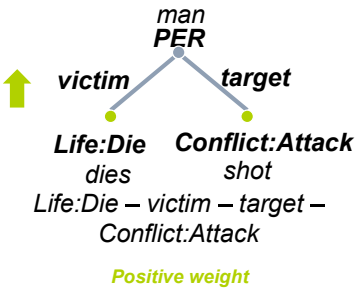
OneIE: An End-to-end Neural Model for IE

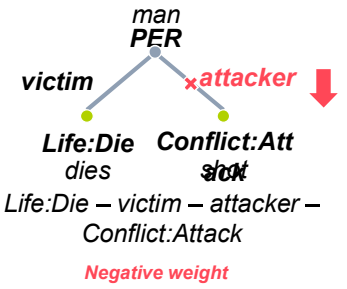


- **Decoding**: In the test phase, we use a beam search decoder to find the information graph with the **highest global score**

Incorporating Global Features

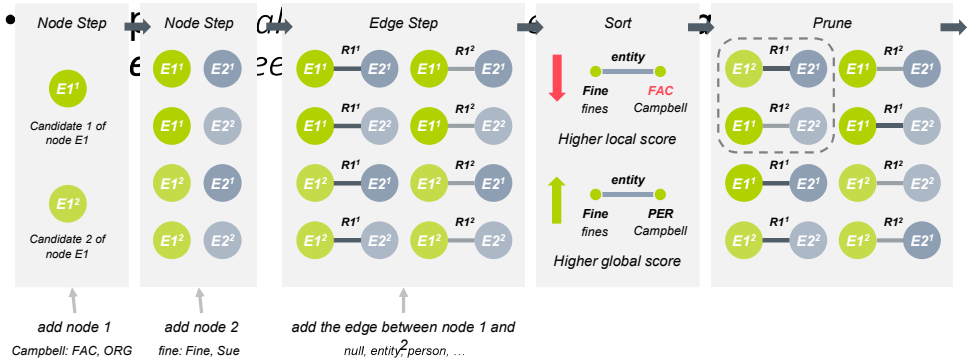
- We design a set of *global feature templates* (e.g., event_type$_1$ – role$_1$ – role$_2$ – event_type$_2$: an entity acts a role$_1$ argument for an event_type$_1$ event and a role$_2$ argument for an event_type$_2$ event in the same sentence)
- The model learns the *weight* of each feature during training

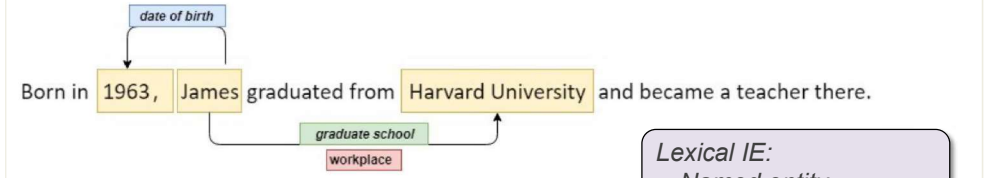




Decoding

- 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=GoNIzi6JtHA
- https://www.youtube.com/watch?v=DnP5uN2EuWA
- https://www.youtube.com/watch?v=8HL-Ap5_Axo
- https://www.youtube.com/watch?v=25u7ZmczdI8
- https://www.youtube.com/watch?v=re5Aw6D7RNo
- https://www.youtube.com/watch?v=PlmNvfyVy_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

Natural Language Processing Applications

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

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

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

innovate achieve lead

- 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

innovate achieve lead

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

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

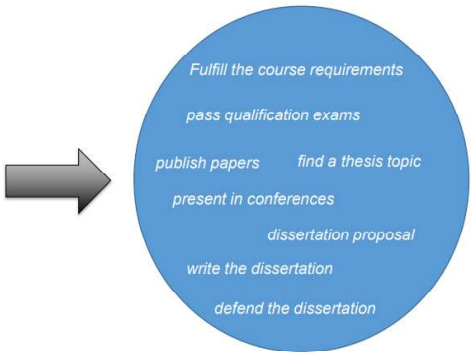
- Features are extracted from the token and its context, and a statistical sequence labeler is trained (any sequence model can be used).

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?

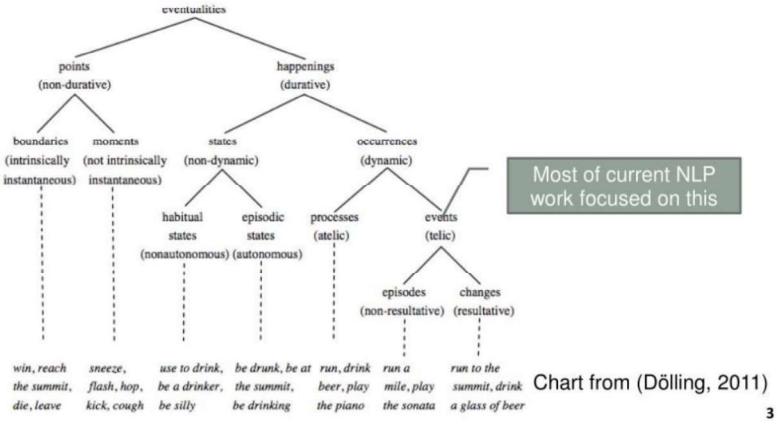
Earning a PhD in Computer Science typically takes around 5 years. It first involves ***fulfilling the course requirements*** and ***passing qualification exams***. Then within several years, the student is expected to ***find a thesis topic***, ***publish several papers*** about the topic and ***present 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.



Most of current NLP work focused on this

Chart from (Dölling, 2011)

3

What is Event?

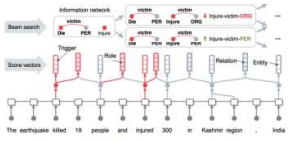
An action or **a series of actions** that happen at a specific location, within **a period of time**, and causes change(s) to the status of **some object(s)**

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.

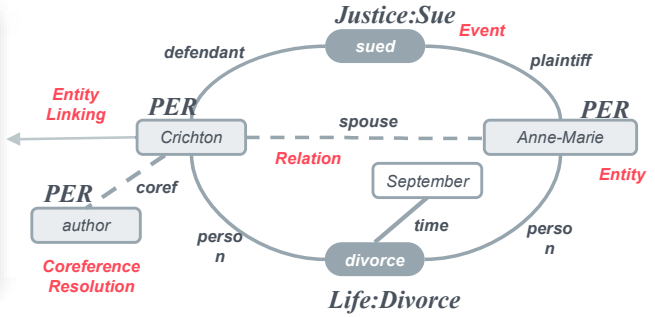
IBM purchased Microsoft.
IBM was purchased by Microsoft.
IBM was purchased on Tuesday by Microsoft

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

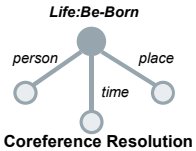
> *Anne-Marie sued Crichton, best known as the author of Jurassic Park, for divorce in September.*



18

Information Extraction Subtasks

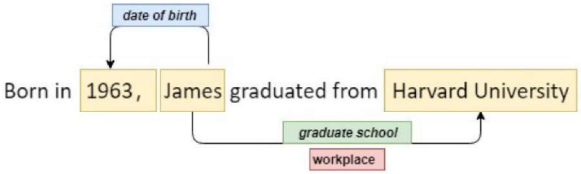
- **Entity Extraction** aims to identify entity mentions in text and classify them into pre-defined entity types.
- **Relation Extraction** is the task of assigning a relation type to an ordered pair of entity mentions.
- **Event Extraction** entails identifying and classifying event triggers and their 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.

19

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. ✗

jealous → angry → bit → scolded ✓

bit → get a cookie ✗

Machine comprehension

Water is split, providing a source of electrons and protons (hydrogen ions, H^+) and giving off O_2 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

Light absorbed → *Water is split* → *Transfer of the electrons and hydrogen 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.

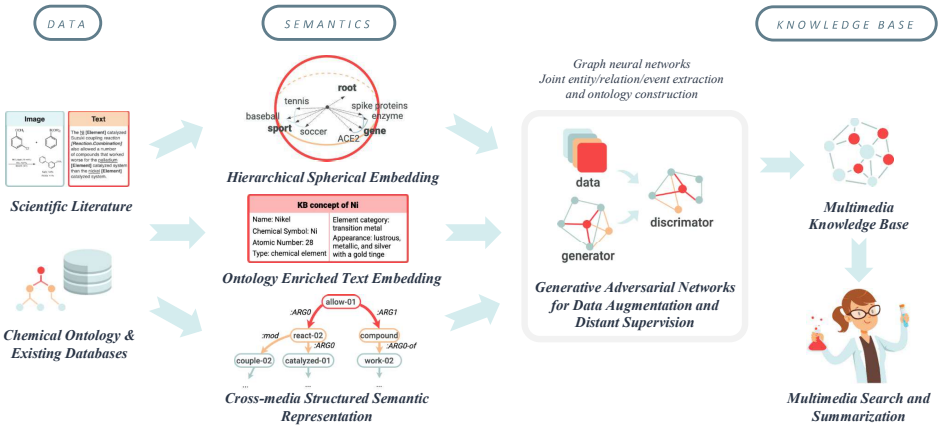
Semi-Structured Text

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



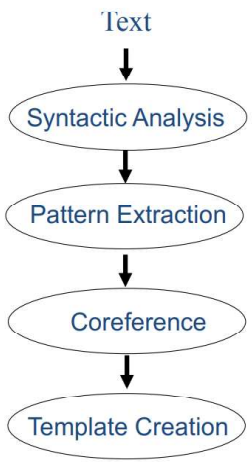
Event mention Extraction

- 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

Argument, role=victim trigger

ACE event type:subtype	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



IBM fired its CEO.

IBM fired its CEO.
Subj *VP* *Dobj*

Event: FIRING
Agent: IBM *Fired:* its CEO

IBM fired its CEO.
John Smith was let go on Monday.

Event: FIRING
Agent: IBM
Fired: John Smith, CEO
Date: Monday

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

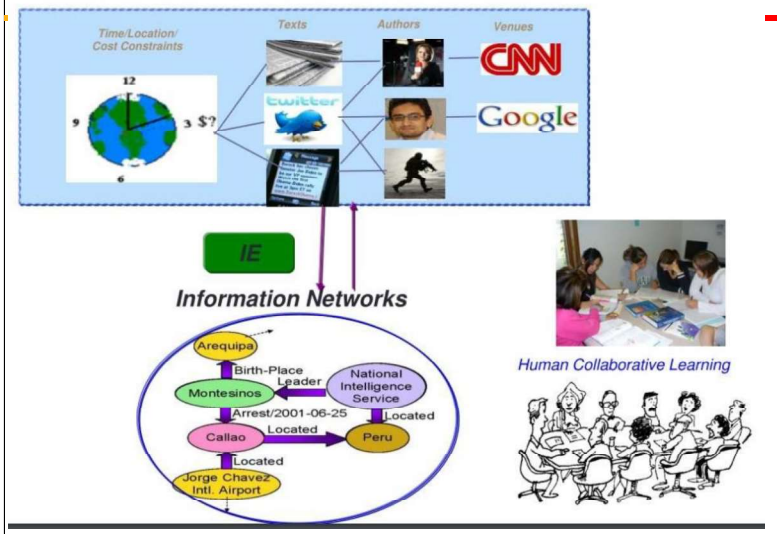
(Ji and Grishman, 2008)

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)

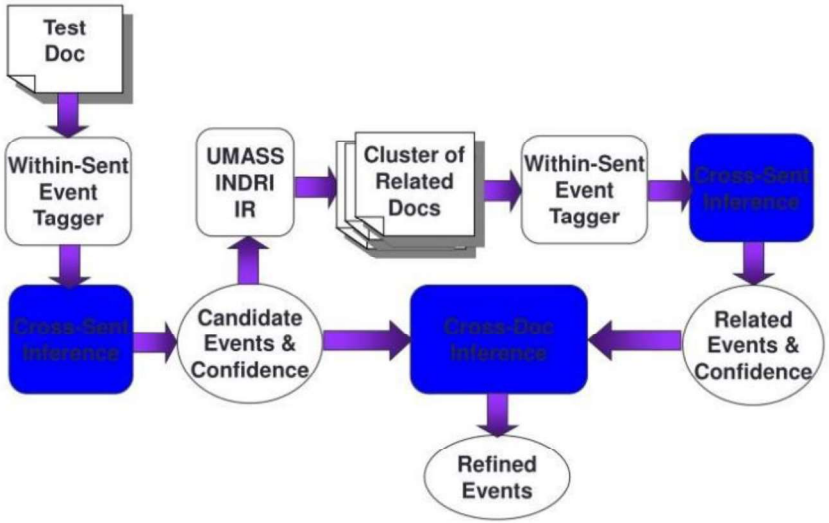
IE in Rich Contexts



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

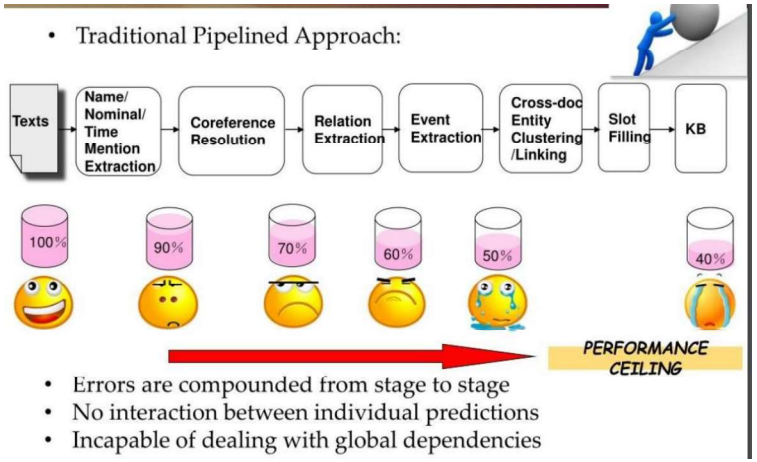
1. Pattern matching
   - Build a pattern from each ACE training example of an event
     - British and US forces reported gains in the advance on Baghdad
       → PER report gain in advance on LOC
2. MaxEnt models
   ① Trigger Classifier
   - to distinguish event instances from non-events, to classify event instances by type

   ② Argument Classifier
   - to distinguish arguments from non-arguments

   ③ Role Classifier
   - to classify arguments by argument role

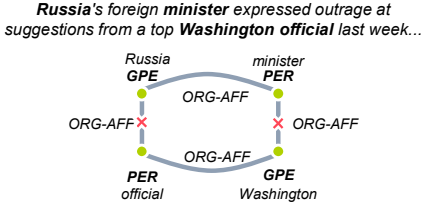
   ④ Reportable-Event Classifier
   - to determine whether there is a reportable event instance

Event Mention Extraction



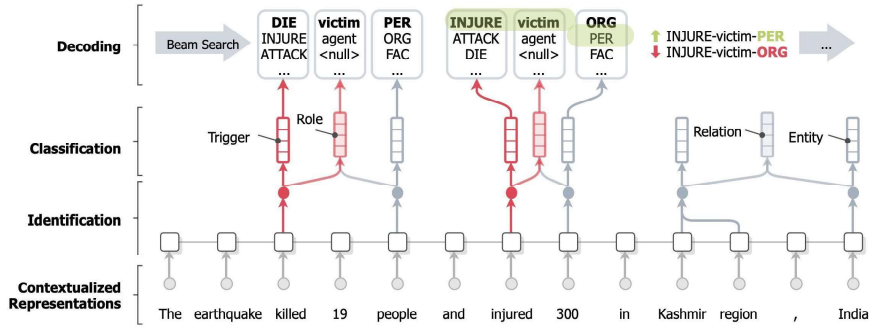
Motivation

- 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.



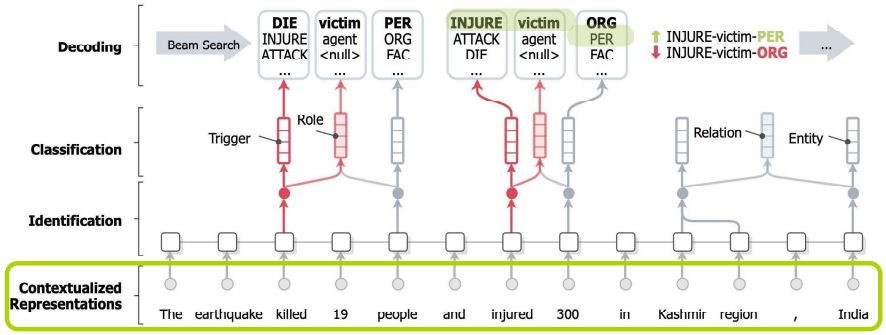
35

OneIE: An End-to-end Neural Model for IE



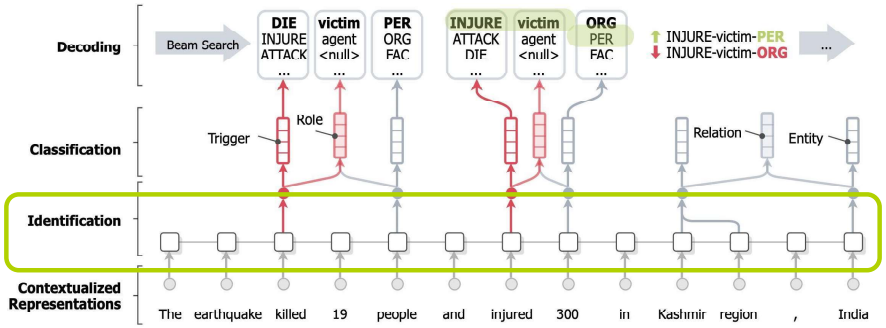
- Our OneIE framework extracts the information graph from a given sentence in four steps: encoding, identification, classification, and decoding

OneIE: An End-to-end Neural Model for IE



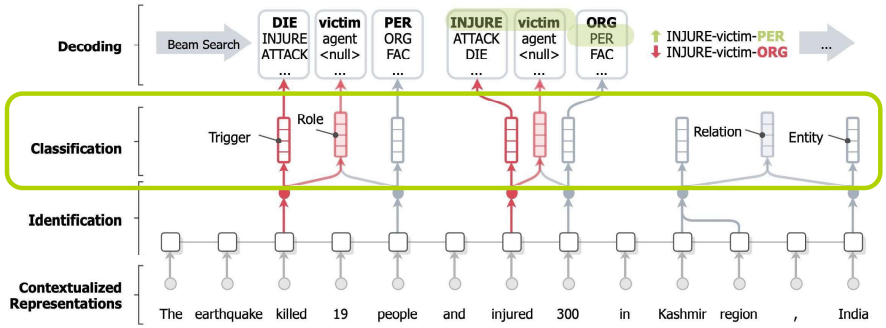
- **Encoding**: We use a BERT encoder to obtain the contextualized embedding of each token

OneIE: An End-to-end Neural Model for IE



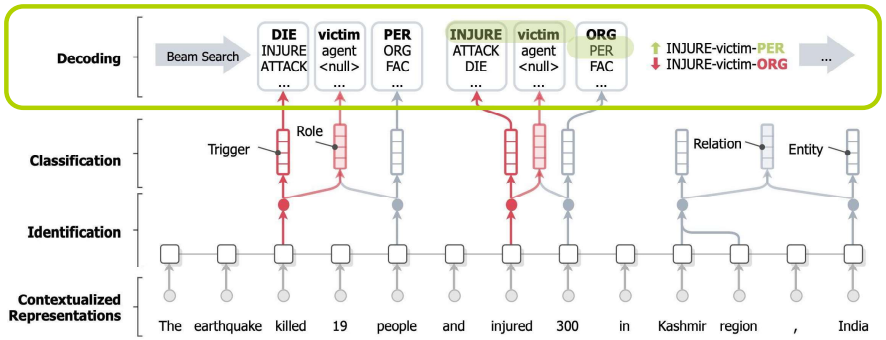
- **Identification**: We use CRF taggers to identify entity mentions and event triggers
- We define the identification loss as $\mathcal{L}^{\mathrm{I}} = -\log p(\boldsymbol{z}|\boldsymbol{X})$

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 $\mathcal{L}^{\mathrm{t}} = -\frac{1}{N^t}\sum_{i=1}^{N^t} \boldsymbol{y}_i^t \log \hat{\boldsymbol{y}}_i^t$

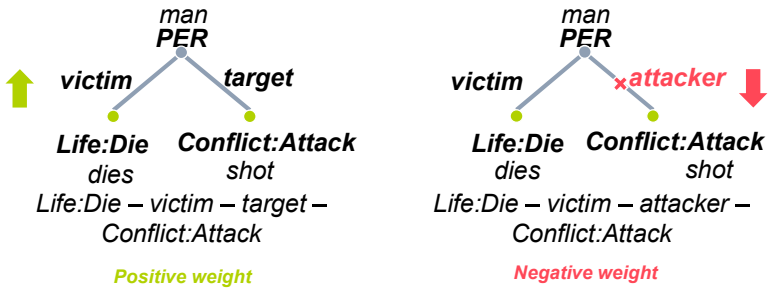
OneIE: An End-to-end Neural Model for IE



- **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., event_type$_1$ – role$_1$ – role$_2$ – 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

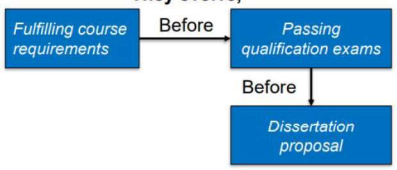
	Features	Weight
1	A **Transport** event has only one **Destination** argument	2.61
2	An **Attack** event has only one **Place** argument	2.31
3	A **PER-SOC** relation exists between two **PER** entities	1.51
4	A **Beneficiary** argument is a **PER** entity	0.93
5	An entity has an **ORG-AFF** relation 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

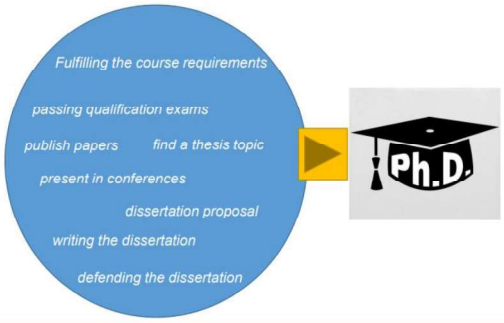
Extraction only is not enough.

Events are **NOT simple, static** predicates.

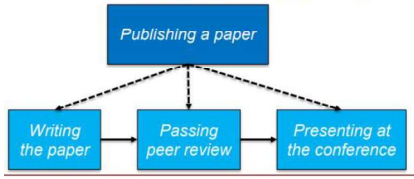
They evolve,



and are always directed by specific intents or central goals [Zacks et al. *Nature Neuroscience*, 2001]



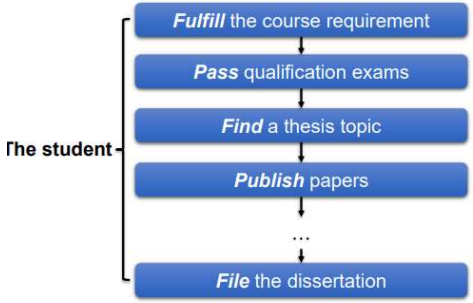
are described in different granularities,



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:

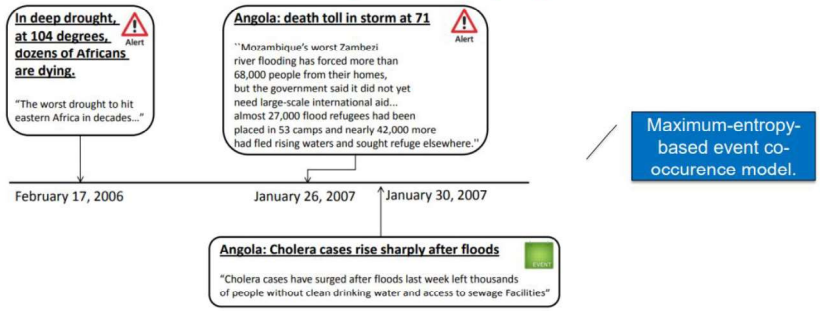
sentenced obj	0.89	indicted obj	0.74
paroled obj	0.76	fined obj	0.73
fired obj	0.75	denied subj	0.73

⇒ Improves narrative cloze tests (36% improvement on NYT Narrative Cloze).

Predicting Events

Radinsky and Horvitz. Mining the Web to Predict Future Events. WSDM, 2013

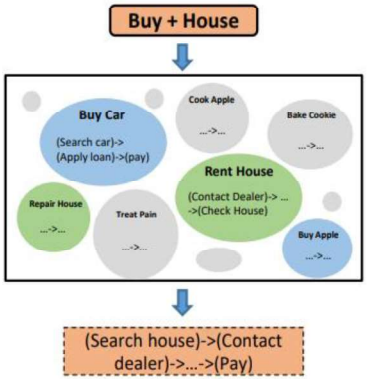
Extension of the event chain model on multiple **dated** and **topically cohesive** documents.



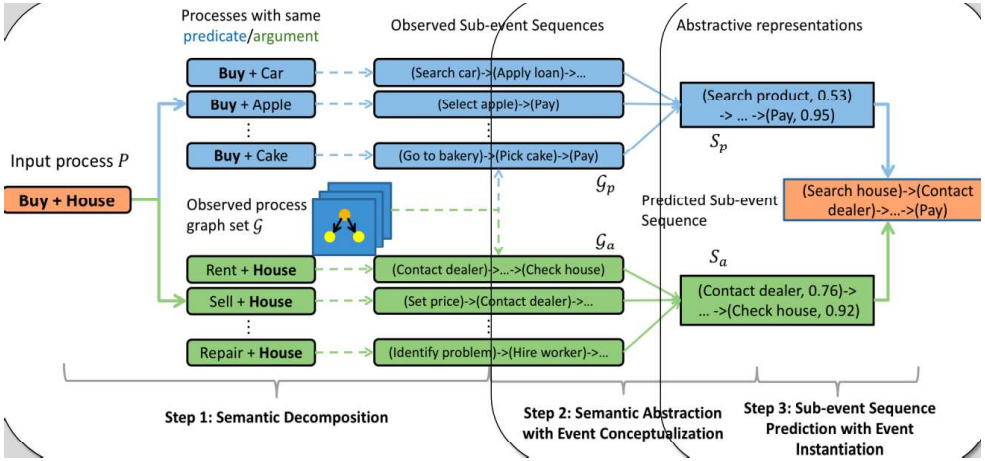
The likelihood of cholera rising is predicted high after a drought followed by storms in Angola (*based on corpus statistics*).

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	to impress their family
	X's reaction	tired, a sense of belonging
	Y's reaction	impressed
PersonX drags PersonX's feet	X's intent	to avoid doing things
	X's reaction	lazy, bored
	Y's reaction	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

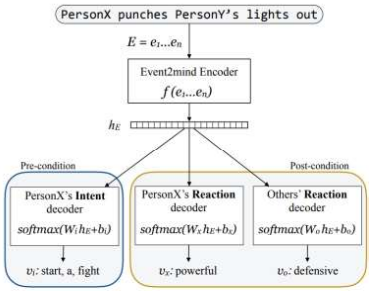
Performs Seq2NGram generation:

PersonX cooks steak

PersonX's intent: ["steak", "to kill their hunger", "to make dinner for the family", "to eat steak"]

PersonX's reaction: ["excited", "accomplished", "proud", "full"]

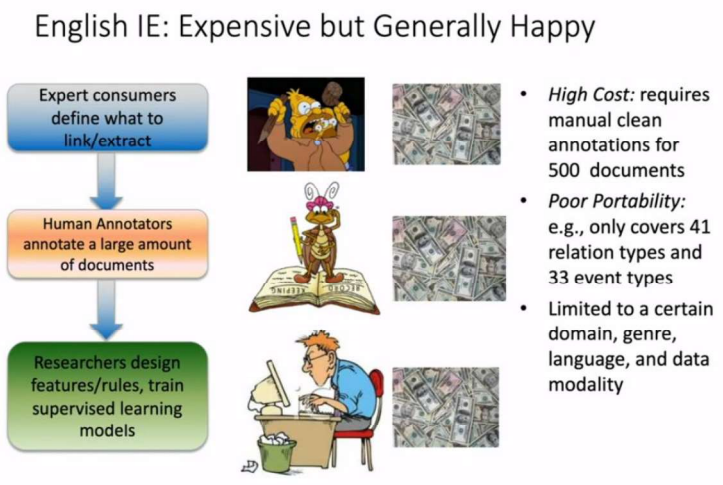
Other people's reaction: ["none", "happy", "person x cooked well."]



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



Cross lingual Event Extraction

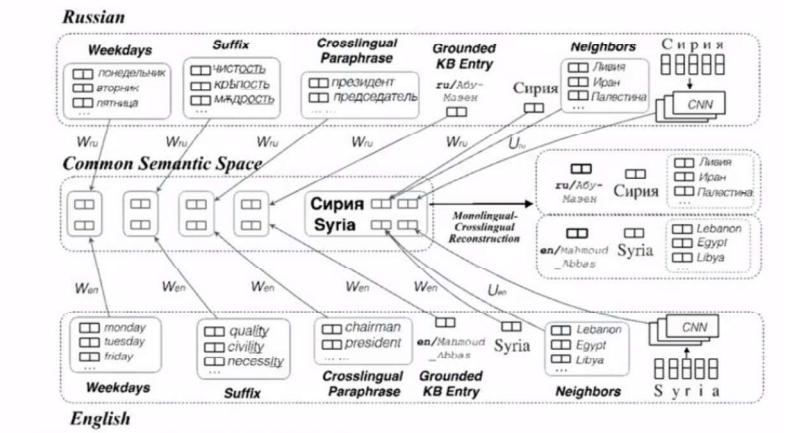
- 3000+ living languages, 300+ languages have digital news data
- Certain information is often reported predominantly in local news in low-resource languages
  - e.g., the vast majority of Physical-Located relations and Meeting events involving Aung San Suu Kyi are only reported locally in Burmese news
  - 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/data modality

Cross lingual Event Extraction



- (Huang et al., EMNLP2018)
- Our new hypothesis: Cluster distribution tends to be consistent across languages (Huang et al., EMNLP2018)

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

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

BITS Pilani
Pilani Campus

BITS Pilani
Pilani Campus

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

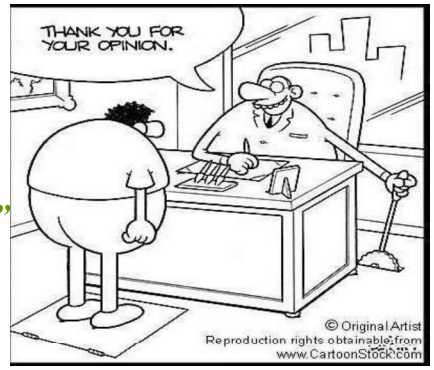
What others think has always been an important piece of information

"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

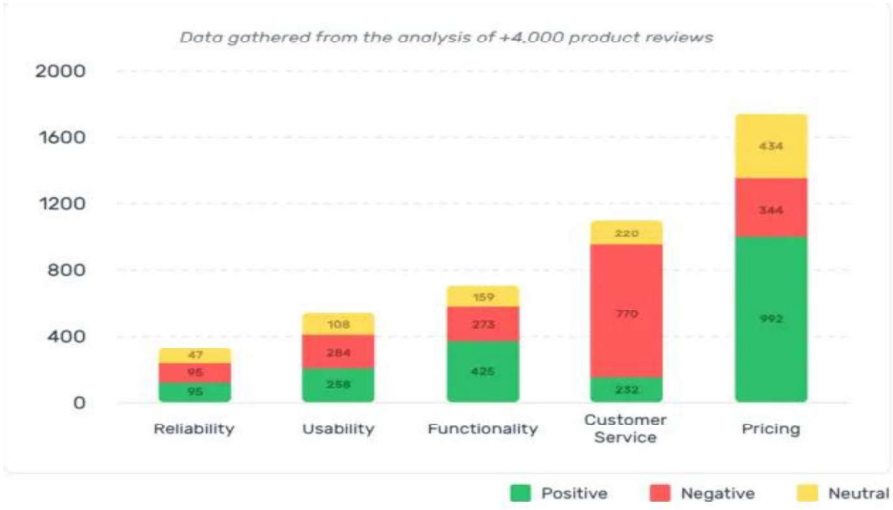
It's a big day & I want to book a table at a nice Japanese restaurant

Sentiment Analysis

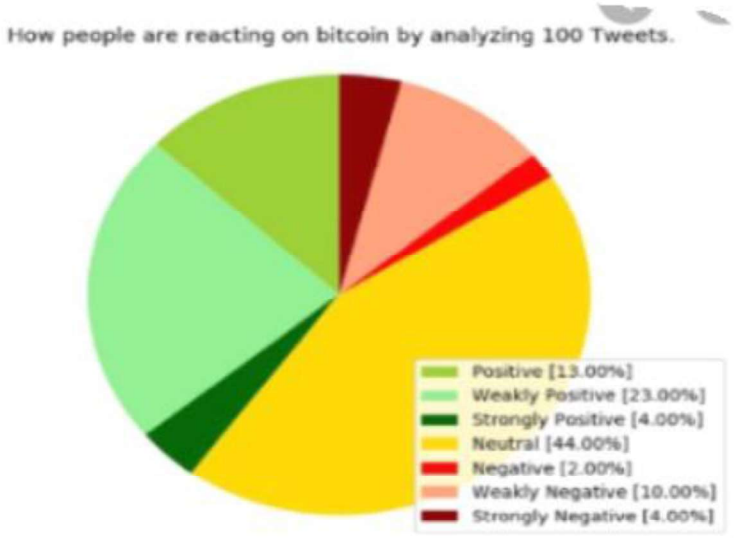
Sentiment Analysis

Examples

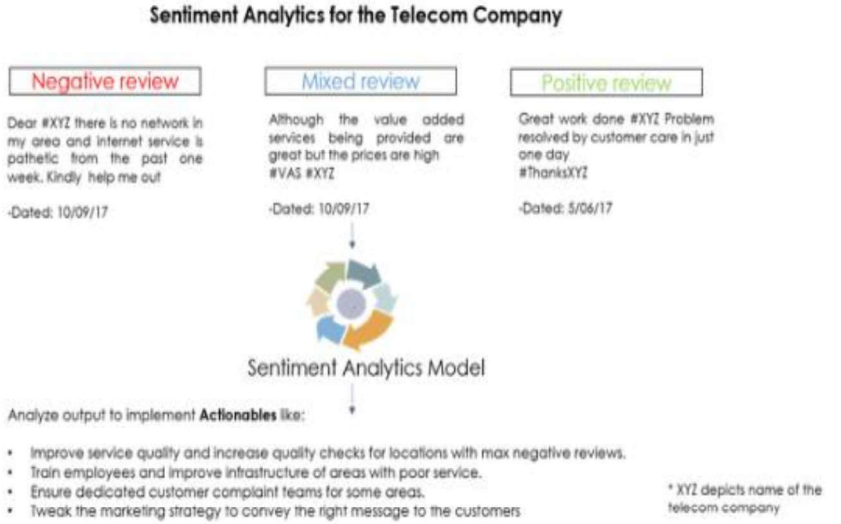
Examples



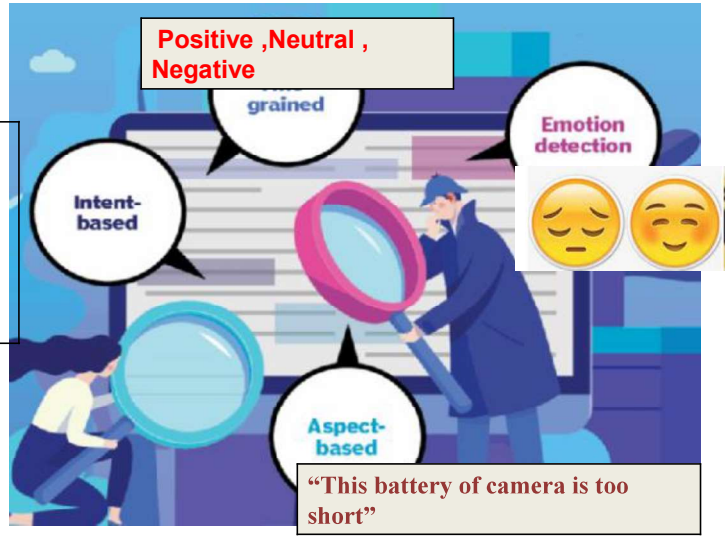
Examples



Examples



Types of Sentiment Analysis

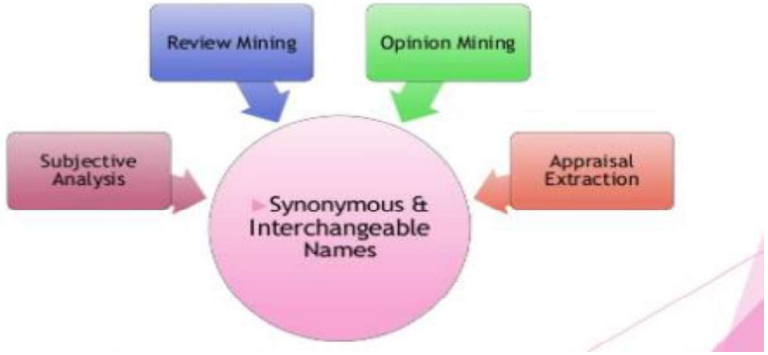


intention to sell, intention to complain or intention to purchase etc

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

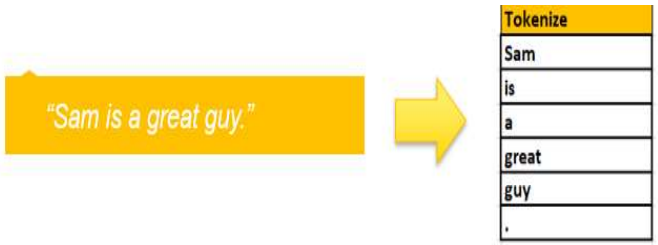
Following steps need to be performed

- 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



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.

Sam	neutral
great	positive
guy	neutral

Sentiment: positive

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..

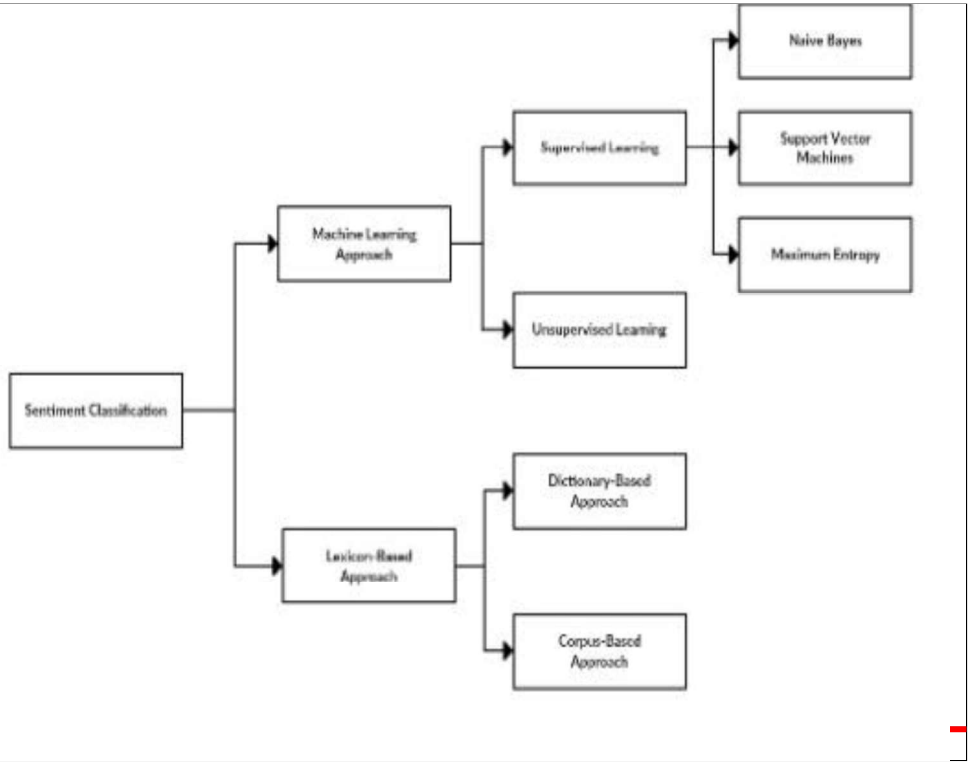
3.Remove stop words.

- Song
- good

4.Classification(Positive, negative ,Neutral)

Apply supervised algorithm

- *Naïve Byes*
- *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	acroscopic#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,tought)
- 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*, asham*, aversi*, avoid*, awkward*
Anger	jealous*, jerk, jerked, jerks, kill*, liar*, lied, lies, lous*, ludicrous*, lying, mad

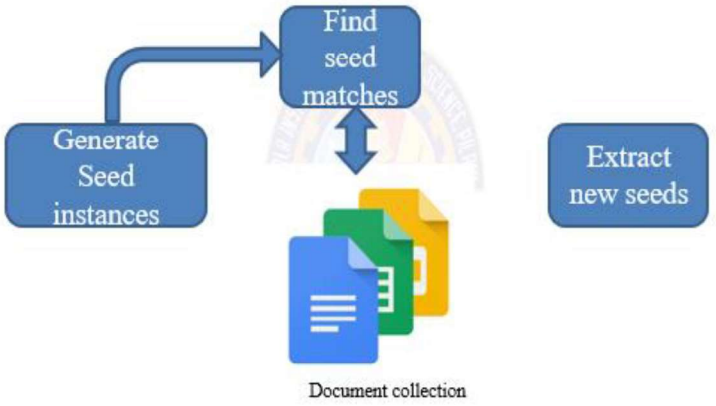
Bing Liu Opinion Lexicon

- 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

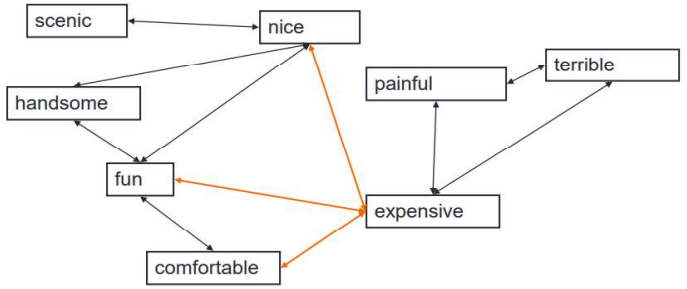
  Fair but brutal

Algorithm

1.Generate a Labeled seed set of adjectives

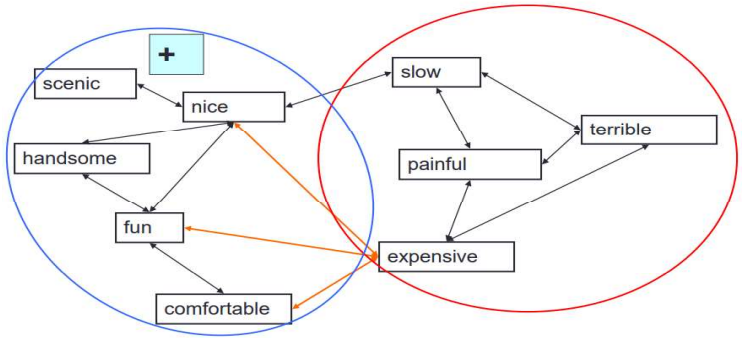
2.Expand seed set to conjoined adjectives by looking up in a corpus/web search

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	Nor 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?

$$\mathrm{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

- If two words are statistically independent, 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”?

–Polarity(phrase) = PMI(phrase, "excellent")– PMI(phrase, "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
Average		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
Average		-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

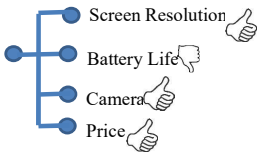
Aspect Based Sentiment Analysis (ABSA)

"(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. … "

.

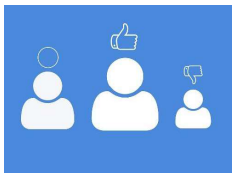
Aspect Based Sentiment Analysis (ABSA)

- Each opinion is a 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

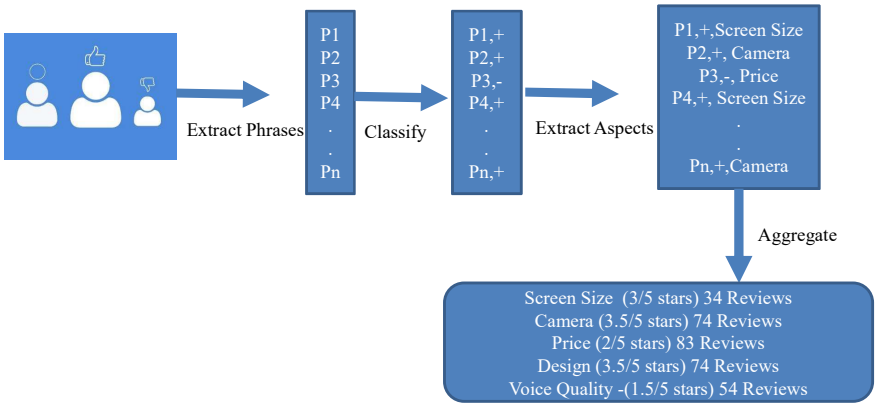
Screen Size – 100/500
Camera Resolution – 300/500
Battery Life – 350/500
Price -450/500
Voice clarity – 325/500

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

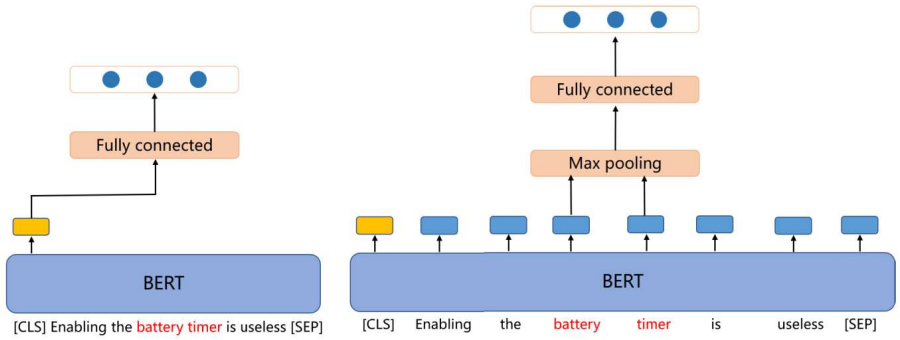


FIGURE 2. The architecture of BERT-FC (left) and TD-BERT (right).

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.

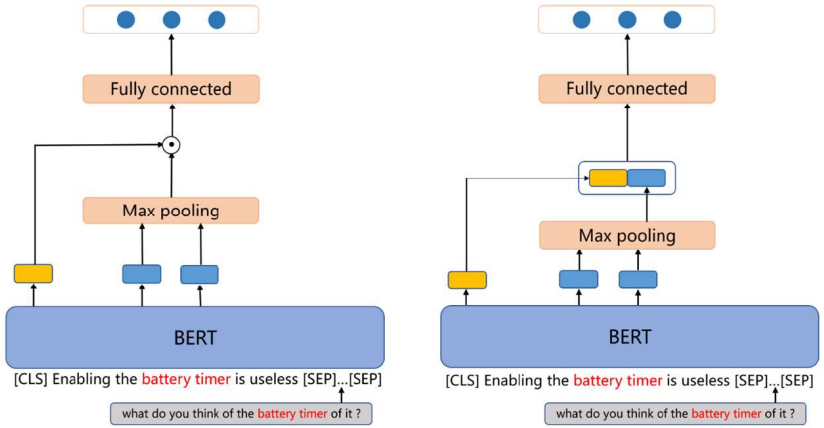


FIGURE 3. The architecture of TD-BERT-QA-MUL (left) and TD-BERT-QA-CON (right).

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 e: pensive, LOCATION1 is often considered th 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 $\mathbf{s = \{w1, \cdots, wm\}}$,

Targets $\mathbf{t = \{t1, \cdots, tk\}}$,

Aspect = {general, price, …}

- Predict sentiment polarity

$\mathbf{y \in \{positive, negative, none\}}$

over $\mathbf{\{(t, a) : t \in T, a \in A\}}$

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

Example: iPhone

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*

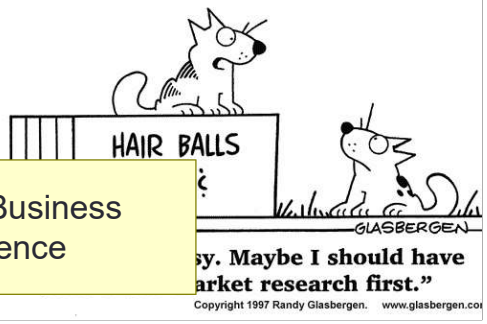
Tech Blog: *The iPhone has a complete set of features*

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)
Ease of Use	(43)
Print Speed	(39)
Connectivity	(31)

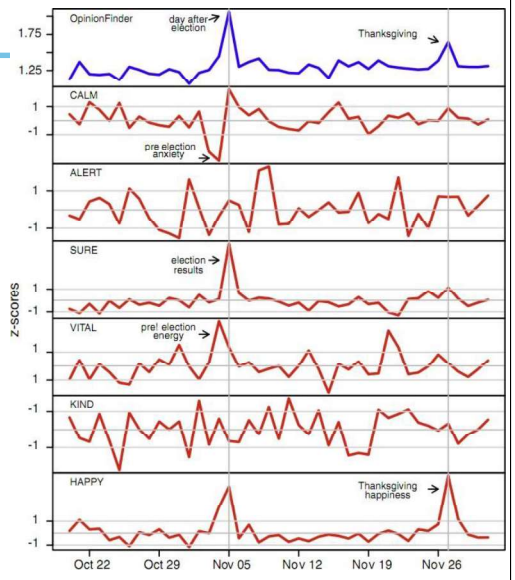
More ▾

Show reviews by source

Best Buy (140)
CNET (5)
Amazon.com (3)

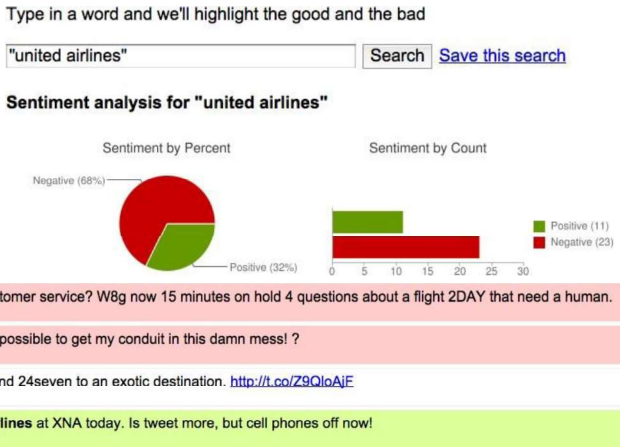
Twitter sentiment:

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

- Twitter Sentiment App
- Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision



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 an 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 critizes 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=OEUlzQawd1s&feature=emb_logo
- VADER Sentiment Analysis: A Complete Guide, Algo Trading and More (quantinsti.com)
- 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/1TK9k41RT8Nf3IhzerNWHpEqWztsk2gAP
- 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-pA5Avcg
- 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