COMP 786 (Fall 2020)
Natural Language Processing

Week 9: Summarization; Guest Talk; Machine Translation 1

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

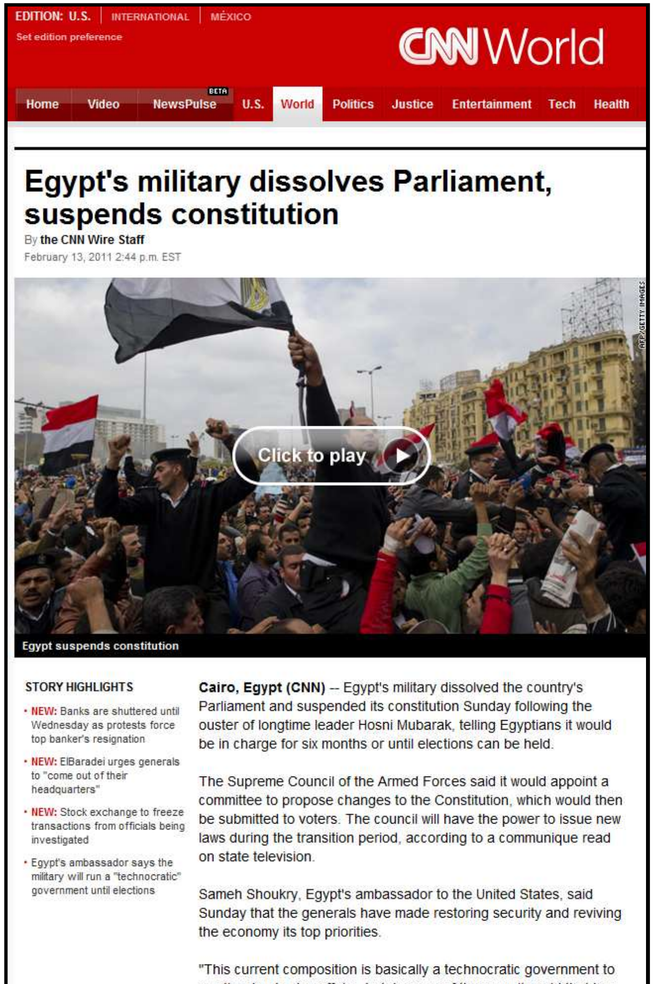
Mohit Bansal

(various slides adapted/borrowed from courses by Dan Klein, JurafskyMartin-SLP3, Manning/Socher, others)

Automatic Document Summarization

Single-Document Summarization

- Full document to a salient, non-redundant summary of ~100 words



EDITION: U.S. | INTERNATIONAL | MÉXICO
Set edition preference

CNN World

Home | Video | NewsPulse BETA | U.S. | World | Politics | Justice | Entertainment | Tech | Health

Egypt's military dissolves Parliament, suspends constitution

By **the CNN Wire Staff**
February 13, 2011 2:44 p.m. EST

Egypt suspends constitution

STORY HIGHLIGHTS

- **NEW:** Banks are shuttered until Wednesday as protests force top banker's resignation
- **NEW:** ElBaradei urges generals to "come out of their headquarters"
- **NEW:** Stock exchange to freeze transactions from officials being investigated
- Egypt's ambassador says the military will run a "technocratic" government until elections

Cairo, Egypt (CNN) -- Egypt's military dissolved the country's Parliament and suspended its constitution Sunday following the ouster of longtime leader Hosni Mubarak, telling Egyptians it would be in charge for six months or until elections can be held.

The Supreme Council of the Armed Forces said it would appoint a committee to propose changes to the Constitution, which would then be submitted to voters. The council will have the power to issue new laws during the transition period, according to a communique read on state television.

Sameh Shoukry, Egypt's ambassador to the United States, said Sunday that the generals have made restoring security and reviving the economy its top priorities.

"This current composition is basically a technocratic government to run the day-to-day affairs, to take care of the security void that has

STORY HIGHLIGHTS

- **NEW:** Banks are shuttered until Wednesday as protests force top banker's resignation
- **NEW:** ElBaradei urges generals to "come out of their headquarters"
- **NEW:** Stock exchange to freeze transactions from officials being investigated
- Egypt's ambassador says the military will run a "technocratic" government until elections

Multi-Document Summarization

- Several news sources with articles on the same topic (can use overlapping info across articles as a good feature for summarization)

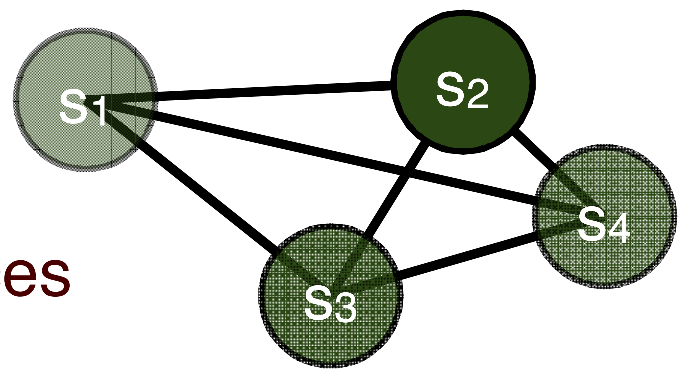
Extractive Summarization

- Directly selecting existing sentences from input document instead of rewriting them

s_1	The health care bill is a major test for the Obama administration.
s_2	Universal health care is a divisive issue.
s_3	President Obama remained calm.
s_4	Obama addressed the House on Tuesday.

Graph-based Extractive Summ

Stationary distribution
represents node centrality



Nodes are sentences

Edges are similarities

[Mihalcea et al., 2004, 2005; inter alia]

Maximize Concept Coverage

s_1	The health care bill is a major test for the Obama administration.
s_2	Universal health care is a divisive issue.
s_3	President Obama remained calm.
s_4	Obama addressed the House on Tuesday.

concept	value
obama	3
health	2
house	1

Length limit: 18 words

summary	length	value	
$\{s_1, s_3\}$	17	5	greedy
$\{s_2, s_3, s_4\}$	17	6	optimal

[Gillick and Favre, 2009]

Maximize Concept Coverage

- A set coverage optimization problem

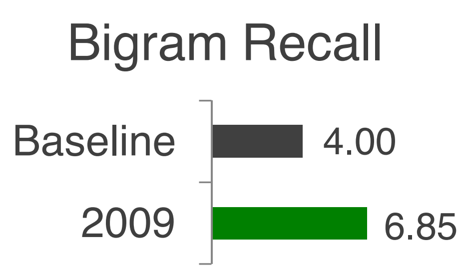
$$\max_{s \in S(D)} \sum_{c \in C(s)} v_c$$

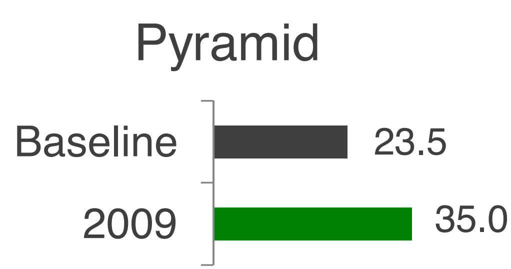
Value of
concept c

Set of extractive summaries
of document set D

Set of concepts
present in summary s

Results





[Gillick and Favre, 2009]

Maximize Concept Coverage

- Can be solved using an integer linear program with constraints:

$$\text{Maximize: } \sum_i w_i c_i$$

total concept value

$$\text{Subject to: } \sum_j l_j s_j \leq L$$

summary length limit

$$s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

maintain consistency between selected sentences and concepts

$$c_i \in \{0, 1\} \quad \forall i$$

$$s_j \in \{0, 1\} \quad \forall j$$

c_i an indicator for the presence of concept i in the summary, and s_j an indicator for the presence of sentence j in the summary. We add Occ_{ij} to indicate the occurrence of concept i in sentence j. *Equations (1) and (2) ensure the logical consistency of the solution: selecting a sentence necessitates selecting all the concepts it contains and selecting a concept is only possible if it is present in at least one selected sentence.*

[Gillick et al., 2008] [Gillick and Favre, 2009]

Beyond Extraction: Compression

- If you had to write a concise summary, making effective use of the 100-word limit, you would remove some information from the lengthy sentences in the original article

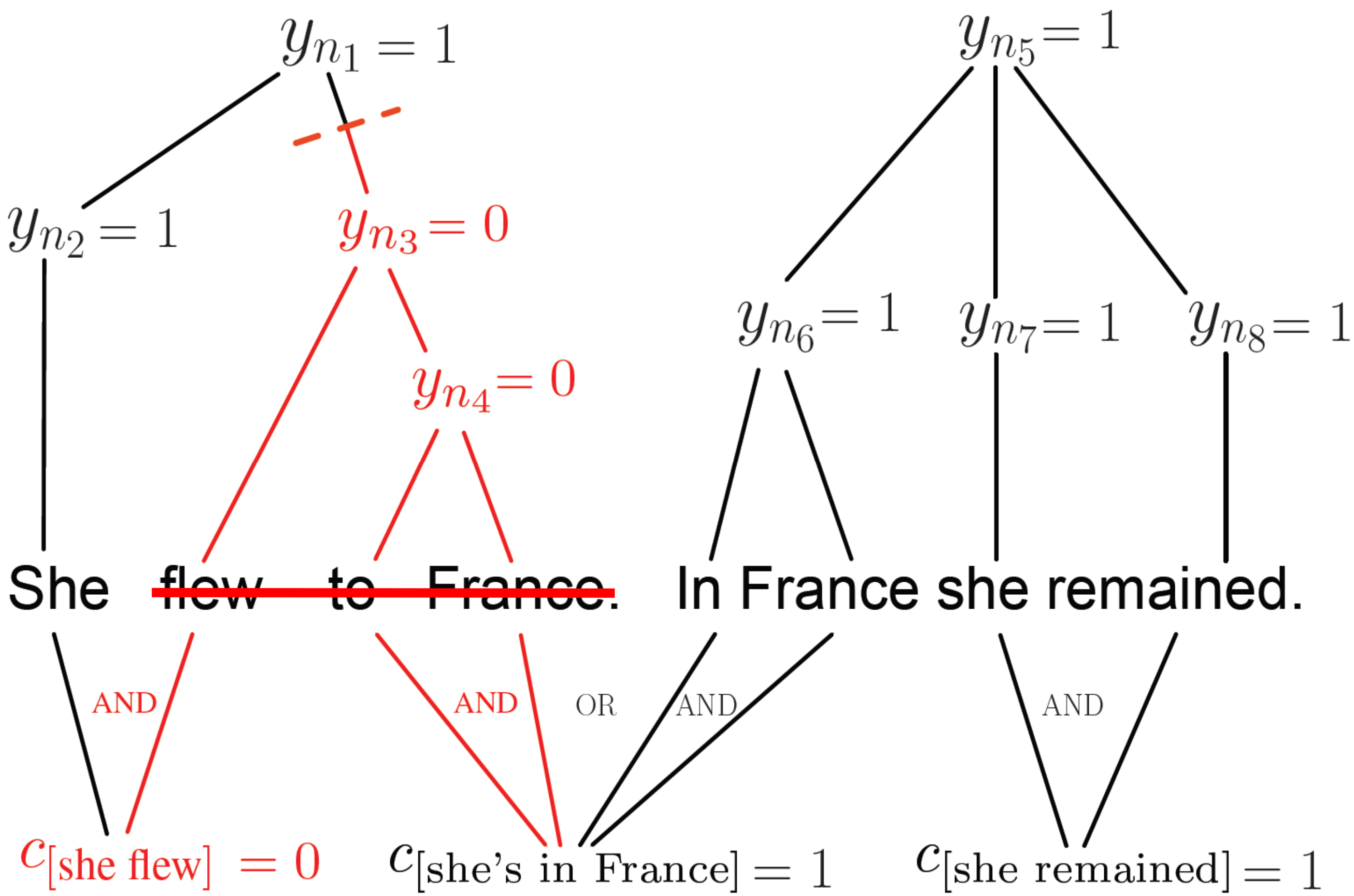
What would a human do?

~~It is therefore unsurprising that~~ Lindsay pleaded not guilty ~~yesterday afternoon~~ to the charges filed against her, ~~according to her publicist.~~

[Berg-Kirkpatrick et al., 2011]

Beyond Extraction: Compression

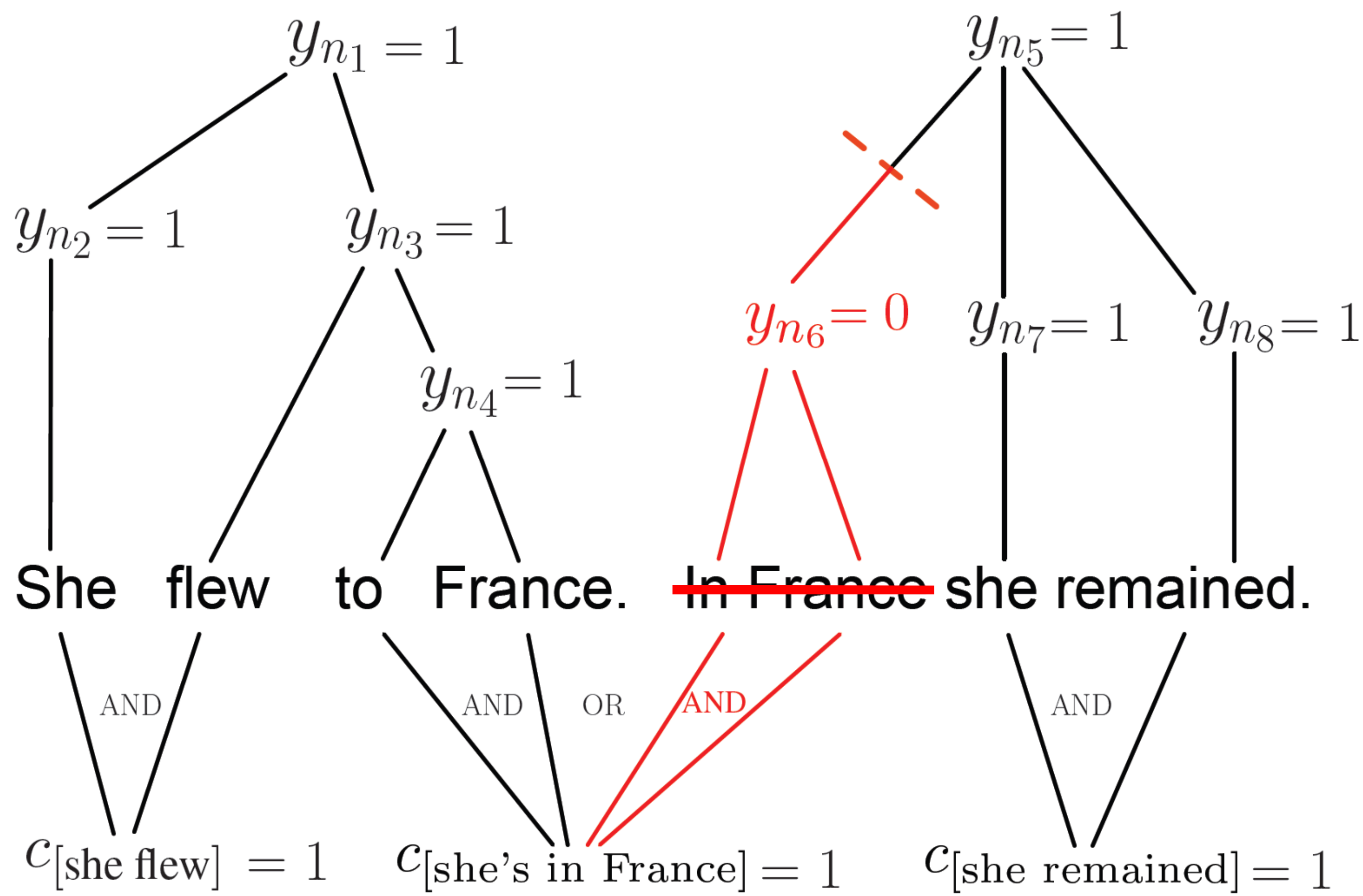
- Model should learn the subtree deletions/cuts that allow compression



[Berg-Kirkpatrick et al., 2011]

Beyond Extraction: Compression

- Model should learn the subtree deletions/cuts that allow compression



[Berg-Kirkpatrick et al., 2011]

Beyond Extraction: Compression

- The new optimization problem looks to maximize the concept values as well as safe deletion values in the candidate summary:

$$\max_{s \in S(D)} \Big[\sum_{c \in C(s)} v_c + \underbrace{\sum_{d \in D(s)}}_{\text{Set branch cut deletions made in creating summary s}} \underbrace{v_d}_{\text{Value of deletion d}} \Big]$$

- To decide the value/cost of a deletion, we decide relevant deletion features and the model learns their weights:

$$v_d = w^\top f(d)$$

[Berg-Kirkpatrick et al., 2011]

Beyond Extraction: Compression

- Some example features for concept bigrams and cuts/deletions:

Bigram Features $f(b)$

COUNT:	Bucketed document counts
STOP:	Stop word indicators
POSITION:	First document position indicators
CONJ:	All two- and three-way conjunctions of above
BIAS:	Always one

Cut Features $f(c)$

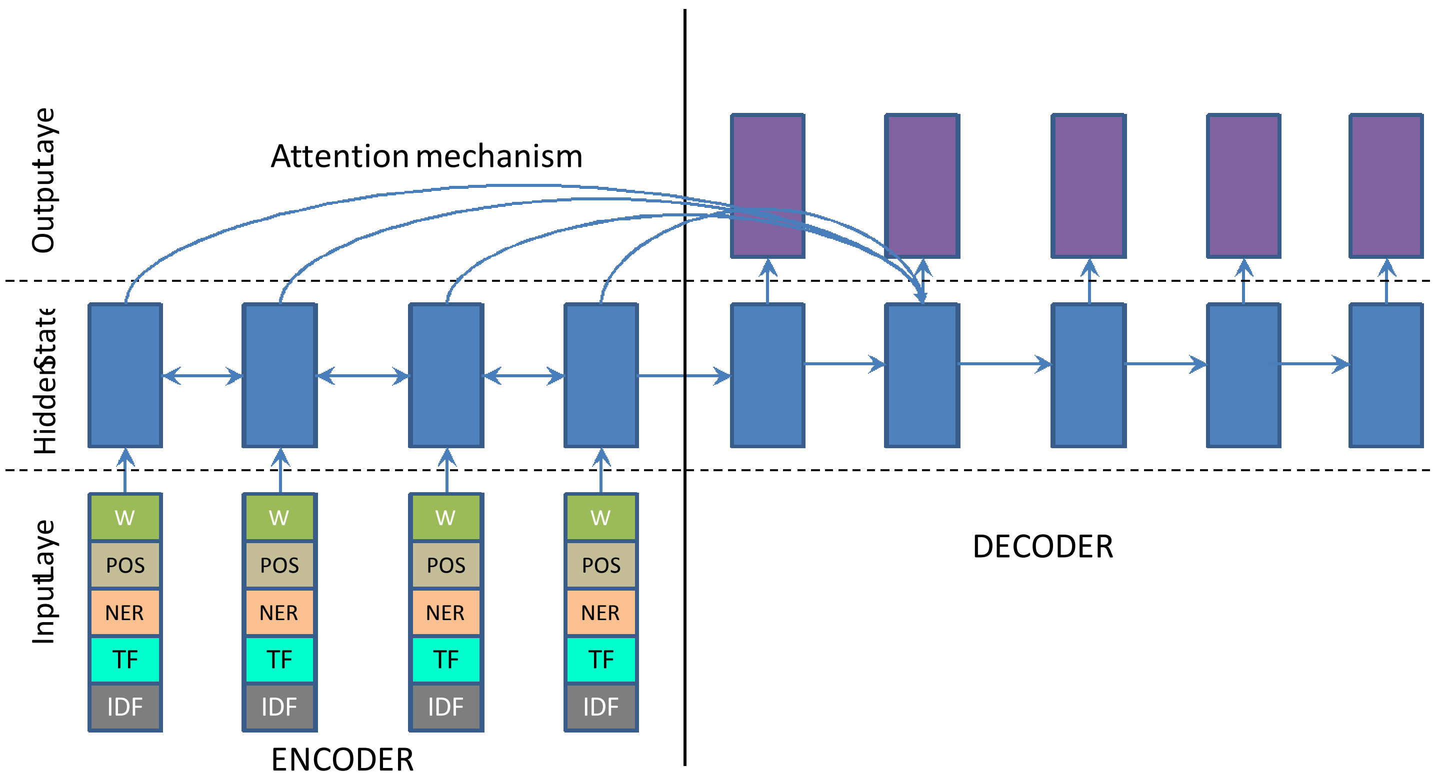
COORD:	Coordinated phrase, four versions: NP, VP, S, SBAR
S-ADJUNCT:	Adjunct to matrix verb, four versions: CC, PP, ADVP, SBAR
REL-C:	Relative clause indicator
ATTR-C:	Attribution clause indicator
ATTR-PP:	PP attribution indicator
TEMP-PP:	Temporal PP indicator
TEMP-NP	Temporal NP indicator
BIAS:	Always one

[Berg-Kirkpatrick et al., 2011]

Neural Abstractive Summarization

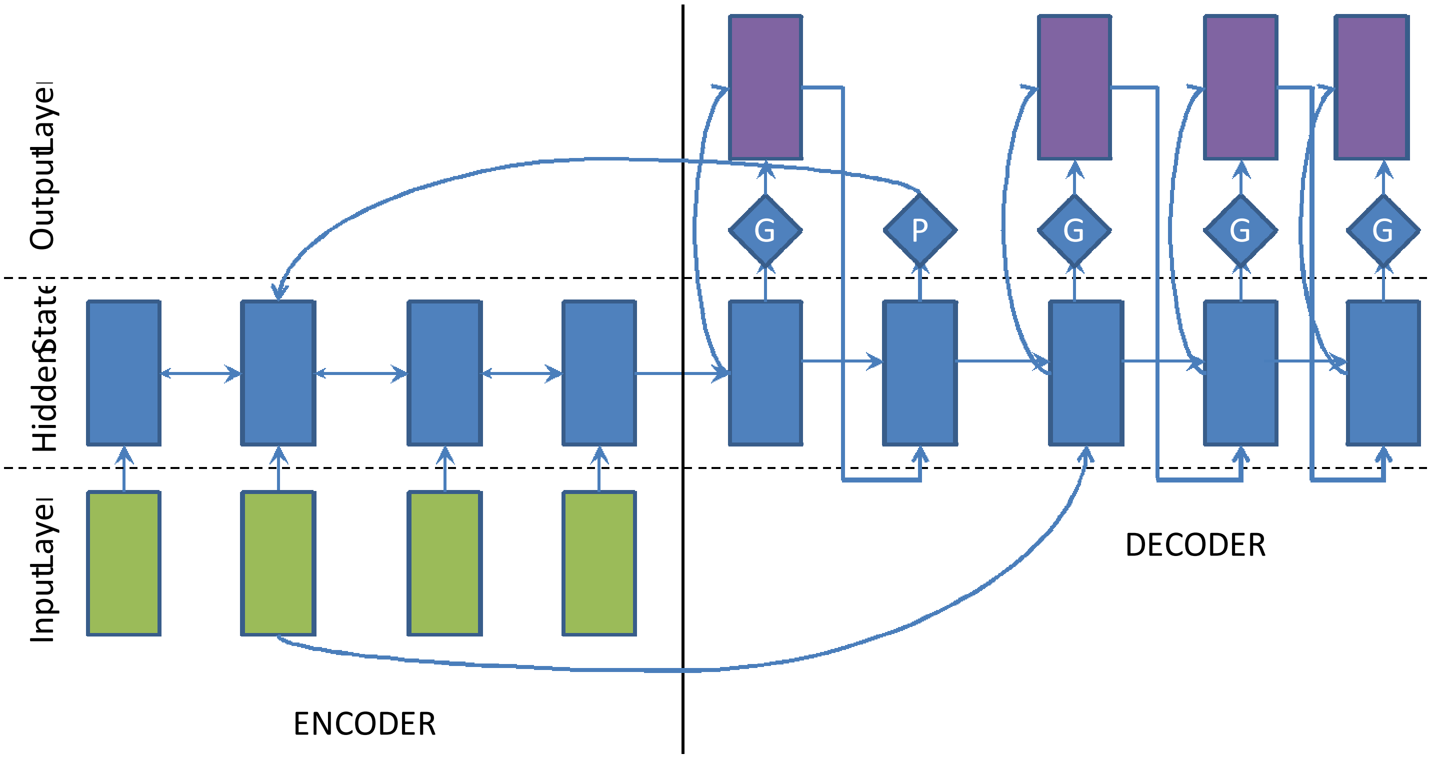
- Mostly based on sequence-to-sequence RNN models
- Later added attention, coverage, pointer/copy, hierarchical encoder/attention, metric rewards RL, etc.
- Examples: Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Paulus et al., 2017

Feature-Augmented Encoder-Decoder



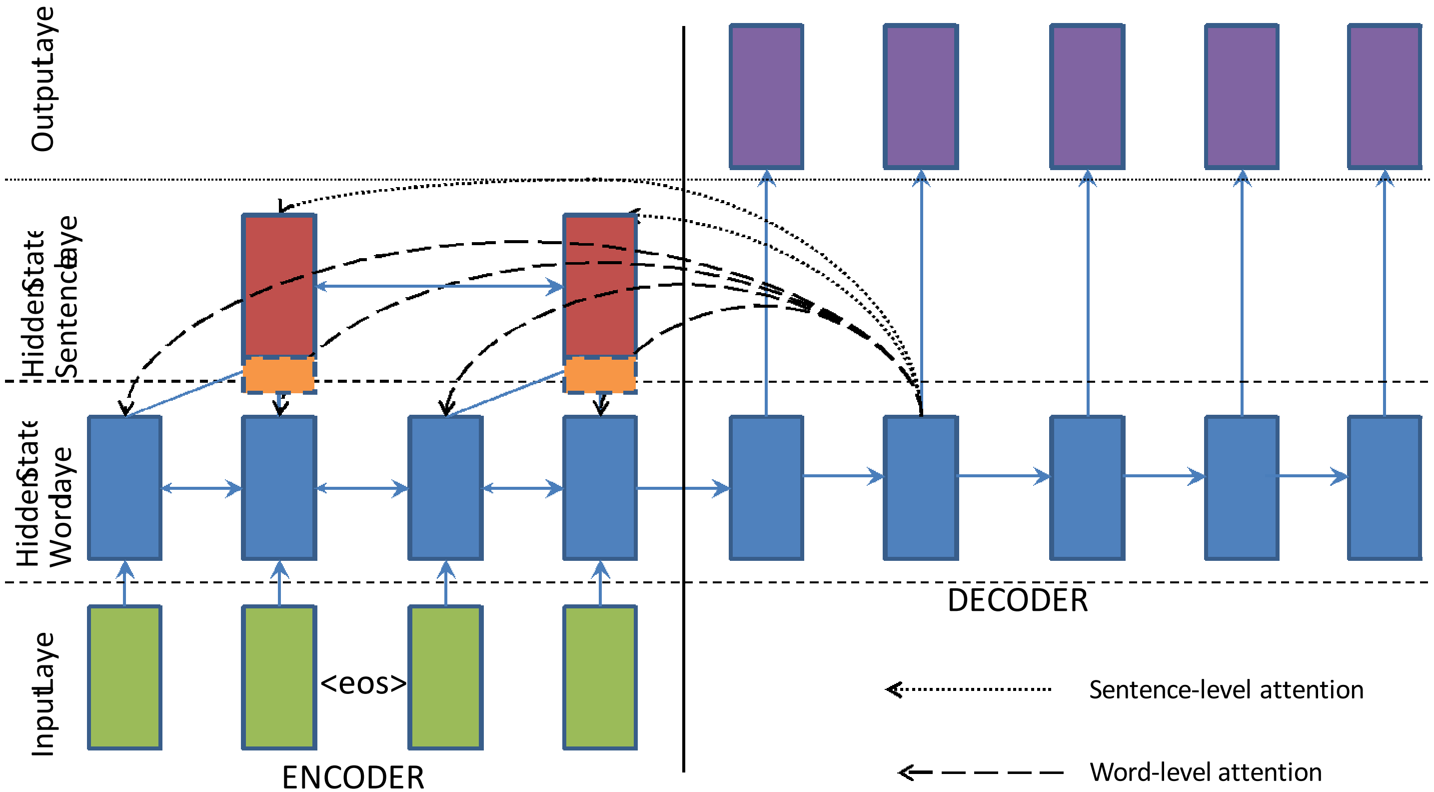
[Nallapati et al., 2016]

Generation+Copying



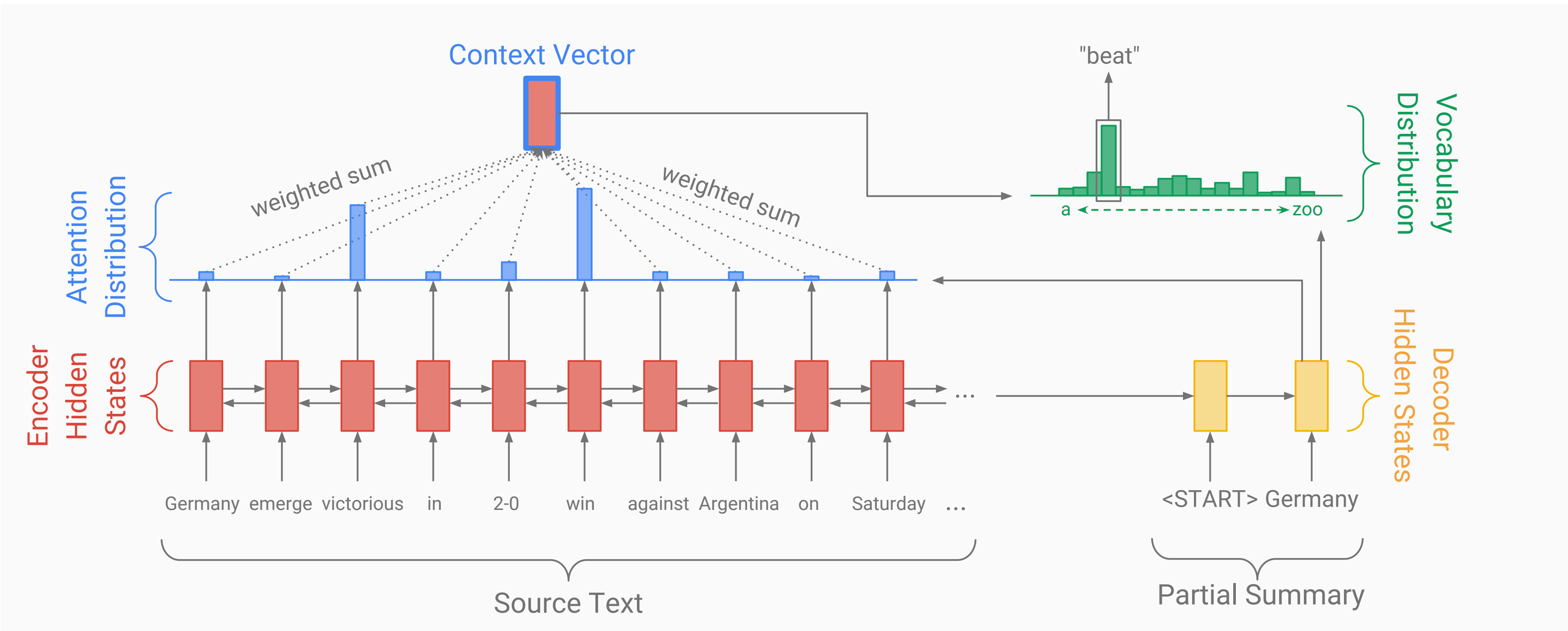
[Nallapati et al., 2016]

Hierarchical Attention



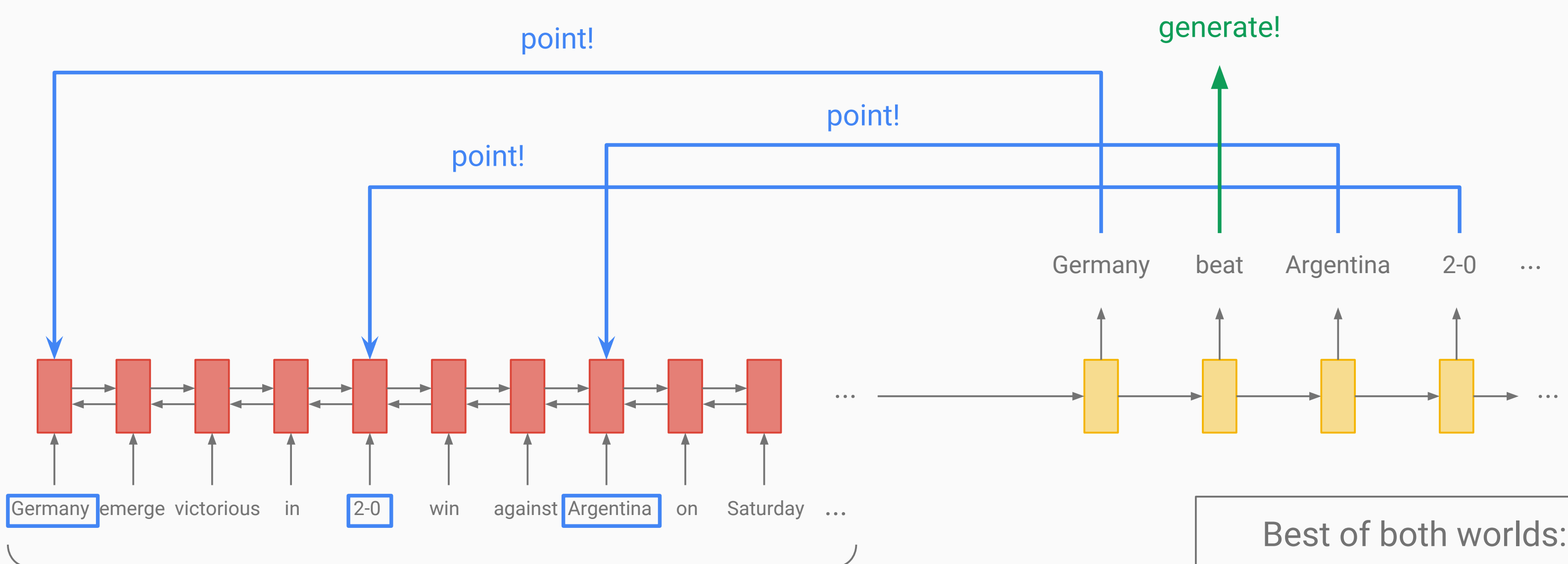
[Nallapati et al., 2016]

Pointer-Generator Networks

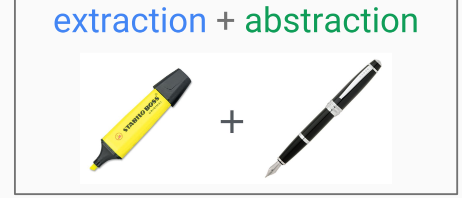


[See et al., 2017]

Pointer-Generator Networks



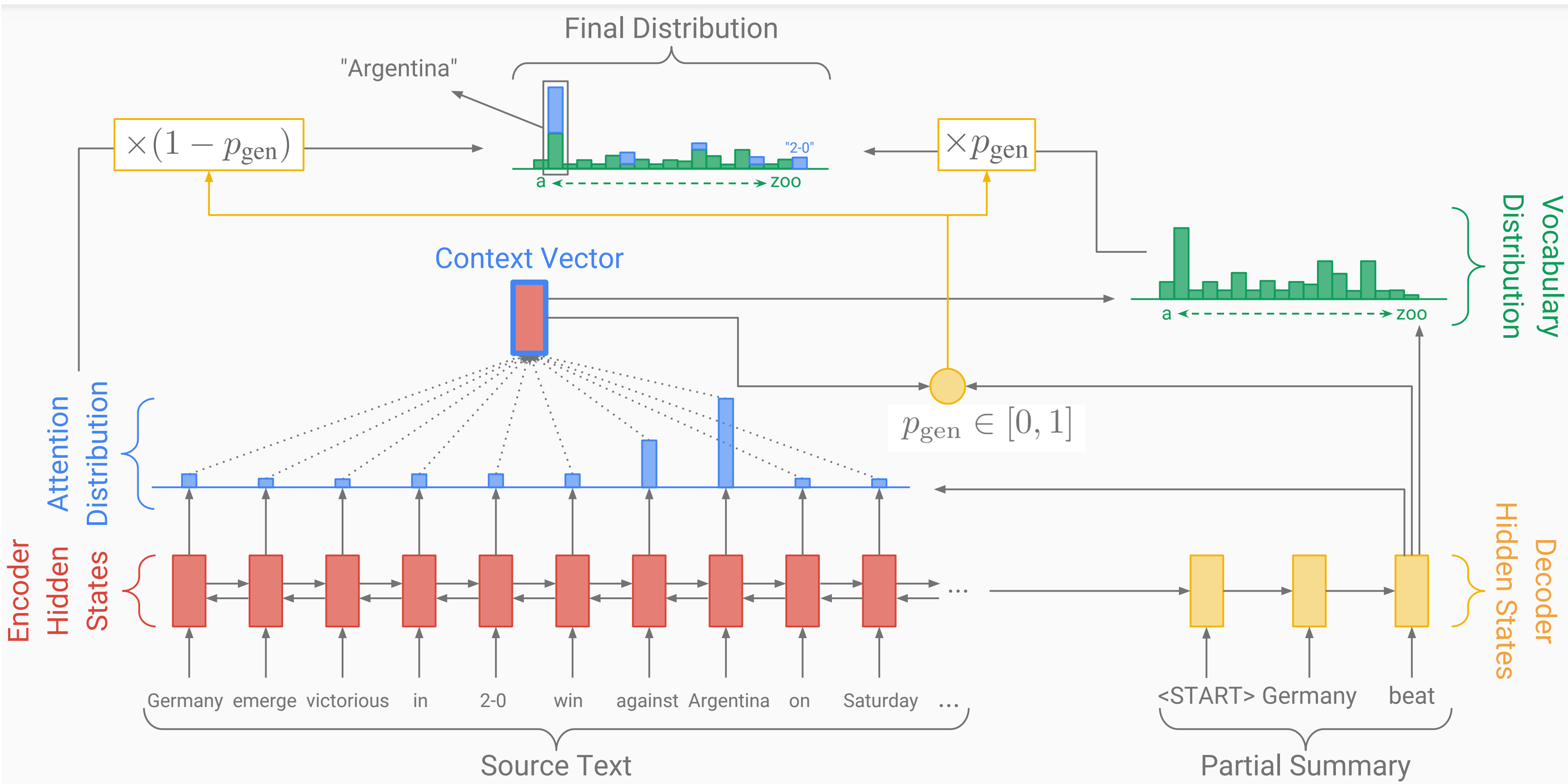
Source Text



*[1] **Incorporating copying mechanism in sequence-to-sequence learning.** Gu et al., 2016.*
*[2] **Language as a latent variable: Discrete generative models for sentence compression.** Miao and Blunsom, 2016.*

[See et al., 2017]

Pointer-Generator Networks



[See et al., 2017]

Coverage for Redundancy Reduction

Coverage = cumulative attention = what has been covered so far

Don't attend here

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday

Summary: Germany beat ______

1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

[4] ***Modeling coverage for neural machine translation.*** *Tu et al., 2016,*
[5] ***Coverage embedding models for neural machine translation.*** *Mi et al., 2016*
[6] ***Distraction-based neural networks for modeling documents.*** *Chen et al., 2016.*

Result: repetition rate reduced to level similar to human summaries

[See et al., 2017]

Guest Talk by Ramakanth Pasunuru:

"Soft, Layer-Specific Multi-Task Summarization with Entailment and Question Generation" (ACL 2018)

"Multi-Reward Reinforced Summarization with Saliency and Entailment" (NAACL 2018)

(20 mins)

Topic:
Abstractive Summarization with Multi-Task Learning and Reinforcement Learning

(presented by Ramakanth Pasunuru)

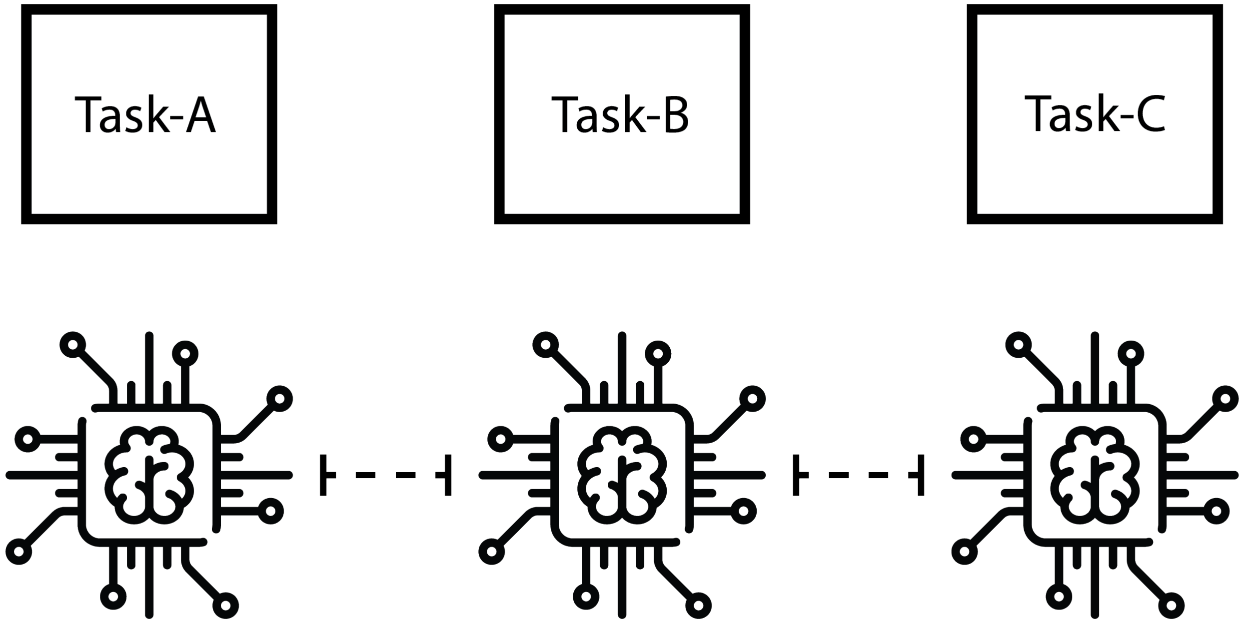
(slides by Ramakanth Pasunuru)

Multi-Task Learning

Multi-Task Learning

- Multi-task Learning (MTL) is an **inductive transfer mechanism** which **leverages information from related tasks** to improve the primary model's **generalization performance**.

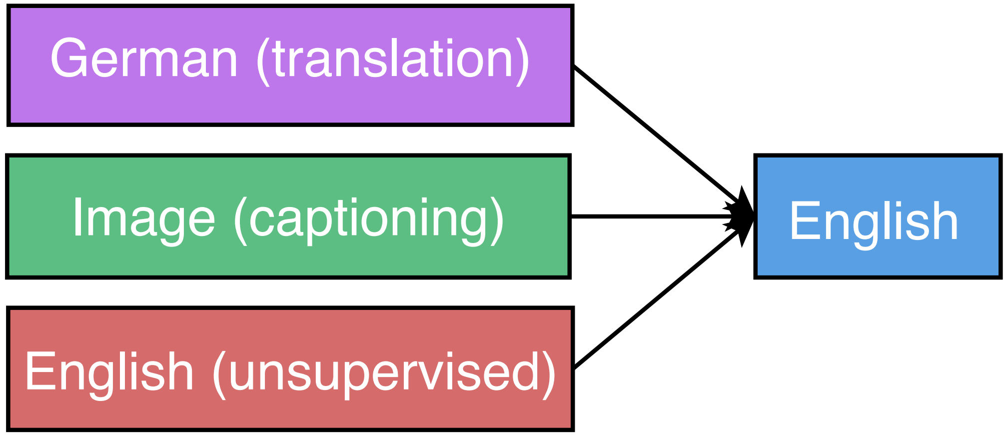
- It achieves this goal by training **multiple** tasks in **parallel** while **sharing representations**, where the training signals from the auxiliary tasks can help improve the performance of the primary task.

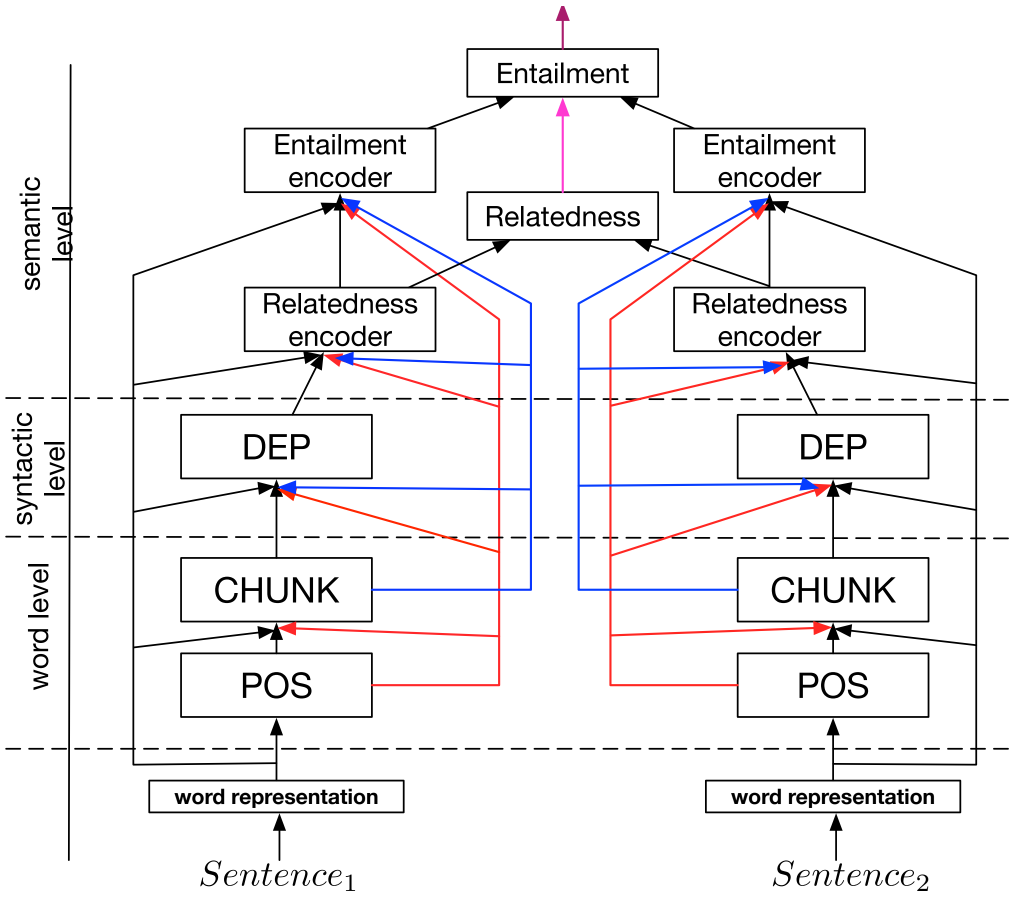


[Caruana, 1998; Argyriou et al., 2007; Kumar and Daume, 2012; Luong et al., 2016]

(slides by Ramakanth Pasunuru)

Previous Work





[Luong et al., 2016]

[Hashimoto et al., 2016]

[Long and Wang, 2015; Lu et al., 2016; Misra et al., 2016; Kendall et al., 2017; Ruder et al., 2018]

(slides by Ramakanth Pasunuru)

MTL for Summarization

MTL for Summarization

- An accurate abstractive summary of a document should contain all its *salient information* and should be *logically entailed* by the input document.

Input Document: *celtic have written to the scottish football association in order to gain an ‘ understanding óf the refereeing decisions during their scottish cup semi-final defeat by inverness on sunday . the hoops were left outraged by referee steven mclean ś failure to award a penalty or red card for a clear handball in the box by josh meekings to deny leigh griffith ś goal-bound shot during the first-half . caley thistle went on to win the game 3-2 after extra-time and denied rory delia ś men the chance to secure a domestic treble this season . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings . celtic ś adam matthews -lrb- right -rrb- slides in with a strong challenge on nick ross in the scottish cup semi-final . ‘ given the level of reaction from our supporters and across football , we are duty bound to seek an understanding of what actually happened , ćeltic said in a statement . they added , ‘ we have not been given any other specific explanation so far and this is simply to understand the circumstances of what went on and why such an obvious error was made . h́owever , the parkhead outfit made a point of congratulating their opponents , who have reached the first-ever scottish cup final in their history , describing caley as a ‘ fantastic club ánd saying ‘ reaching the final is a great achievement . ćeltic had taken the lead in the semi-final through defender virgil van dijk ś curling free-kick on 18 minutes , but were unable to double that lead thanks to the meekings controversy . it allowed inverness a route back into the game and celtic had goalkeeper craig gordon sent off after the restart for scything down marley watkins in the area . greg tansey duly converted the resulting penalty . edward ofere then put caley thistle ahead , only for john guidetti to draw level for the bhoys . with the game seemingly heading for penalties , david raven scored the winner on 117 minutes , breaking thousands of celtic hearts . celtic captain scott brown -lrb- left -rrb- protests to referee steven mclean but the handball goes unpunished . griffiths shows off his acrobatic skills during celtic ś eventual surprise defeat by inverness . celtic pair aleksandar tonev -lrb- left -rrb- and john guidetti look dejected as their hopes of a domestic treble end .*

Ground-truth: *celtic were defeated 3-2 after extra-time in the scottish cup semi-final . leigh griffiths had a goal-bound shot blocked by a clear handball. however, no action was taken against offender josh meekings. the hoops have written the sfa for an ’understanding’ of the decision .*

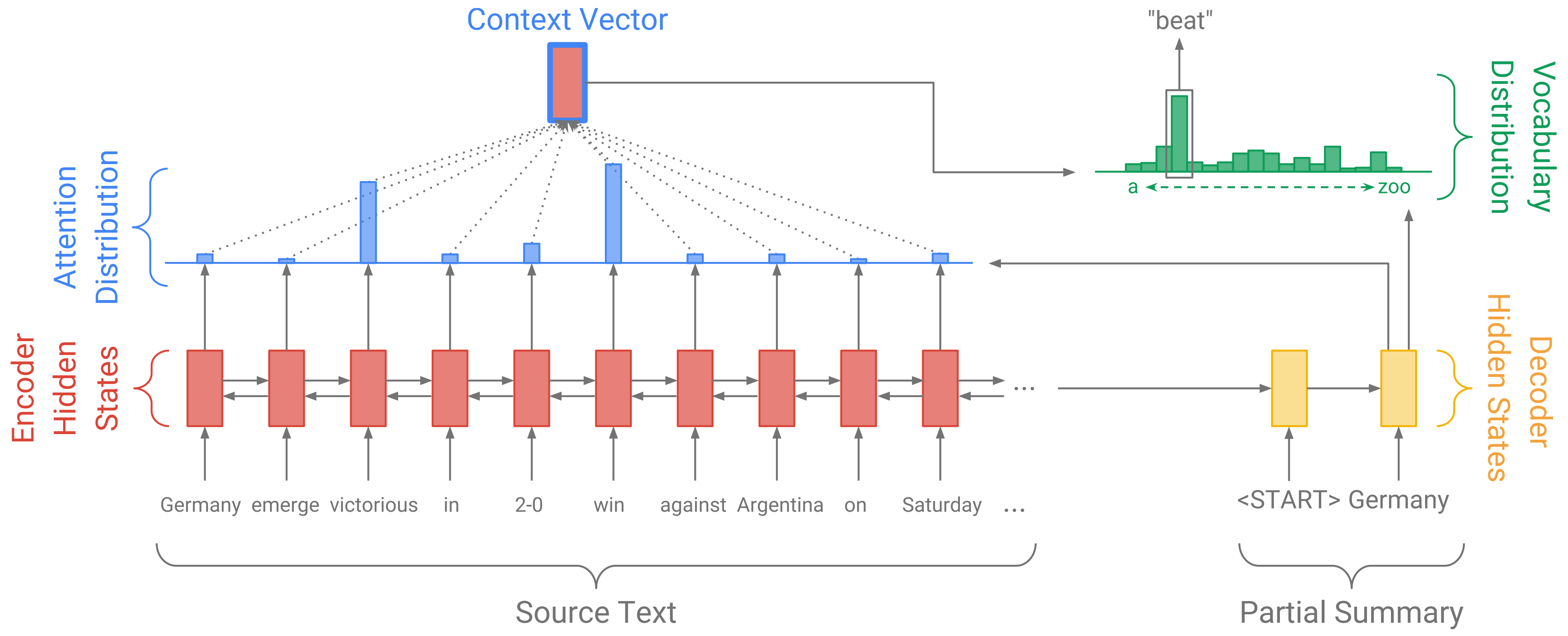
See et al. (2017): *john hartson was once on the end of a major hampden injustice while playing for celtic . but he can not see any point in his old club writing to the scottish football association over the latest controversy at the national stadium . hartson had a goal wrongly disallowed for offside while celtic were leading 1-0 at the time but went on to lose 3-2 .*

MTL for Summarization

- An accurate abstractive summary of a document should contain all its *salient information* and should be *logically entailed* by the input document.
- We improve these via multi-task learning with auxiliary tasks of question generation and entailment generation.
- Question Generation teaches the summarization model how to look for salient questioning-worthy details.
- Entailment Generation teaches the model how to rewrite a summary which is a directed-logical subset of the input document.

Input Document: *celtic have written to the scottish football association in order to gain an ' understanding óf the refereeing decisions during their scottish cup semi-final defeat by inverness on sunday . the hoops were left outraged by referee steven mclean ś failure to award a penalty or red card for a clear handball in the box by josh meekings to deny leigh griffith ś goal-bound shot during the first-half . caley thistle went on to win the game 3-2 after extra-time and denied rory delia ś men the chance to secure a domestic treble this season . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings . celtic ś adam matthews -lrb- right -rrb- slides in with a strong challenge on nick ross in the scottish cup semi-final . ' given the level of reaction from our supporters and across football , we are duty bound to seek an understanding of what actually happened , ćeltic said in a statement . they added , ' we have not been given any other specific explanation so far and this is simply to understand the circumstances of what went on and why such an obvious error was made . h́owever , the parkhead outfit made a point of congratulating their opponents , who have reached the first-ever scottish cup final in their history , describing caley as a ' fantastic club ánd saying ' reaching the final is a great achievement . ćeltic had taken the lead in the semi-final through defender virgil van dijk ś curling free-kick on 18 minutes , but were unable to double that lead thanks to the meekings controversy . it allowed inverness a route back into the game and celtic had goalkeeper craig gordon sent off after the restart for scything down marley watkins in the area . greg tansey duly converted the resulting penalty . edward ofere then put caley thistle ahead , only for john guidetti to draw level for the bhoys . with the game seemingly heading for penalties , david raven scored the winner on 117 minutes , breaking thousands of celtic hearts . celtic captain scott brown -lrb- left -rrb- protests to referee steven mclean but the handball goes unpunished . griffiths shows off his acrobatic skills during celtic ś eventual surprise defeat by inverness . celtic pair aleksandar tonev -lrb- left -rrb- and john guidetti look dejected as their hopes of a domestic treble end .*
Ground-truth: *celtic were defeated 3-2 after extra-time in the scottish cup semi-final . leigh griffiths had a goal-bound shot blocked by a clear handball. however, no action was taken against offender josh meekings. the hoops have written the sfa for an 'understanding' of the decision .*
See et al. (2017): *john hartson was once on the end of a major hampden injustice while playing for celtic . but he can not see any point in his old club writing to the scottish football association over the latest controversy at the national stadium . hartson had a goal wrongly disallowed for offside while celtic were leading 1-0 at the time but went on to lose 3-2 .*

Summarization Model



[See et al., 2017]

Auxiliary Task: Question Generation

- The task of question generation is to generate a question from a given input sentence, which in turn is related to the skill of being able to find the important salient information to ask questions about the sentence.
- A good summary should also be able to find and extract all the salient information in the given source document, and hence we incorporate such capabilities into our abstractive text summarization model by multi-task learning it with a question generation task, sharing some common parameters/representations.

Sentence:
Oxygen is used in cellular respiration and released by photosynthesis, which uses the energy of sunlight to produce oxygen from water.

Questions:
– What life process produces oxygen in the presence of light?
photosynthesis
– Photosynthesis uses which energy to form oxygen from water?
sunlight
– From what does photosynthesis get oxygen?
water

Image Credits: Du et al., 2017

[Rajpurkar et al., 2016; Du et al., 2017]

Auxiliary Task: Entailment Generation

- Directional, logical-implication relation between two sentences:
  - **Premise:** *A girl is jumping on skateboard in the middle of a red bridge.*
  - Entailment: *The girl does a skateboarding trick.*
  - Contradiction: *The girl skates down the sidewalk.*
  - Neutral: *The girl is wearing safety equipment.*

  - **Premise:** *A blond woman is drinking from a public fountain.*
  - Entailment: *The woman is drinking water.*
  - Contradiction: *The woman is drinking coffee.*
  - Neutral: *The woman is very thirsty.*

Auxiliary Task: Entailment Generation

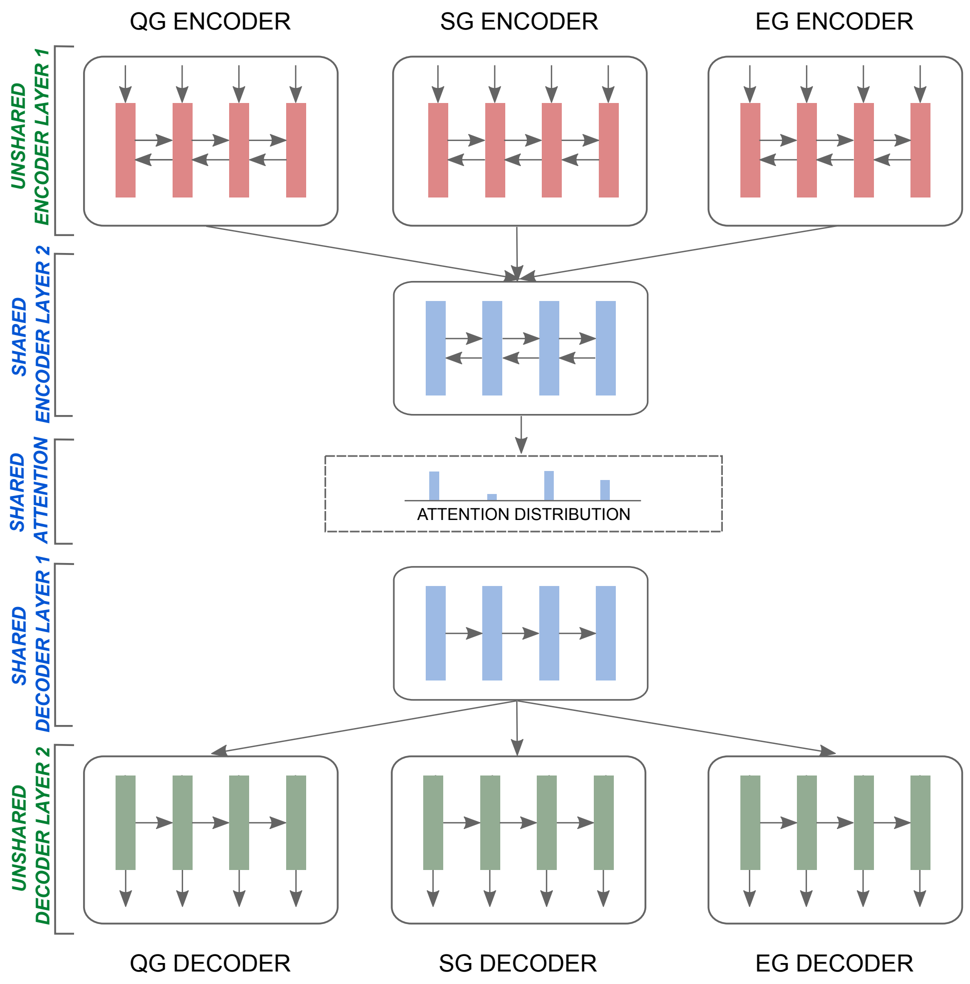
- Directional, logical-implication relation between two sentences:
    - **Premise:** *A girl is jumping on skateboard in the middle of a red bridge.*
    - Entailment: *The girl does a skateboarding trick.*
    - Contradiction: *The girl skates down the sidewalk.*
    - Neutral: *The girl is wearing safety equipment.*

    - **Premise:** *A blond woman is drinking from a public fountain.*
    - Entailment: *The woman is drinking water.*
    - Contradiction: *The woman is drinking coffee.*
    - Neutral: *The woman is very thirsty.*

- The task of entailment generation is to generate a hypothesis which is entailed by (or logically follows from) the given premise as input.

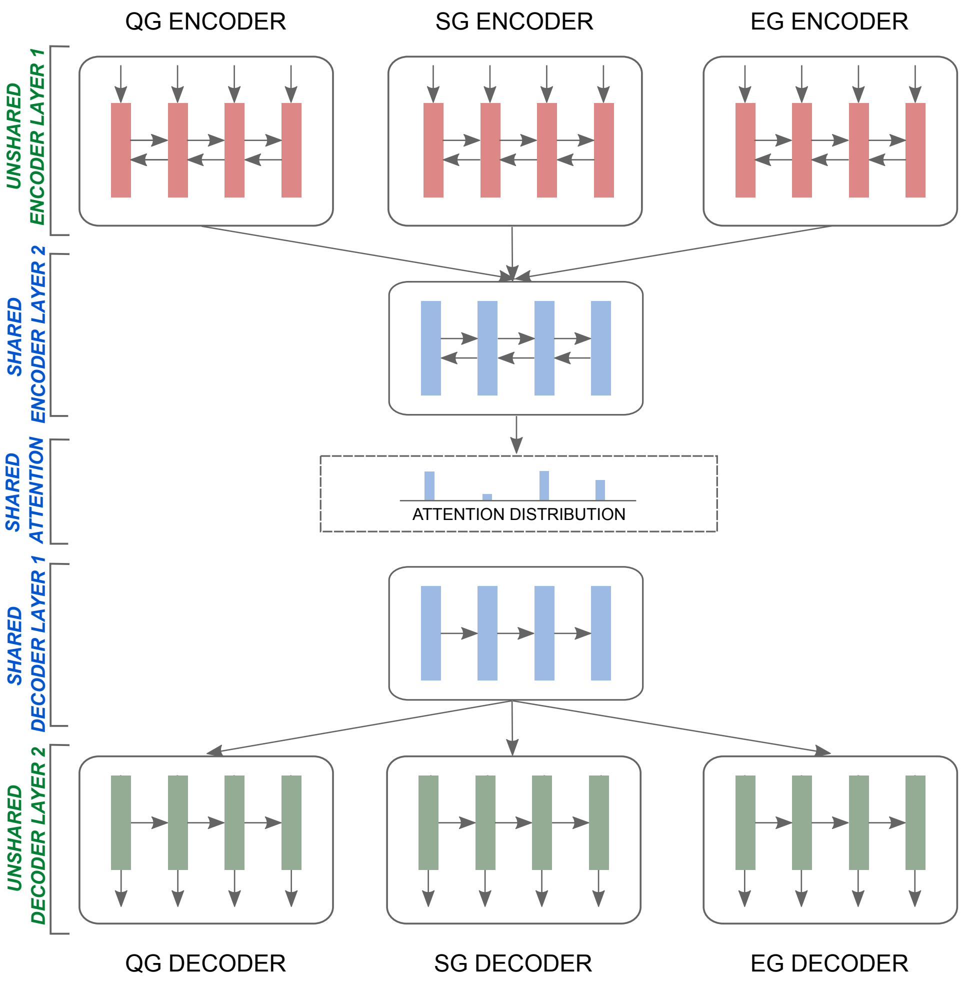
- In summarization, the generation decoder also needs to generate a summary that is entailed by the source document, i.e., does not contain any contradictory or unrelated/extraneous information as compared to the input document.

MTL Architecture



- QG stands for Question Generation
- SG stands for Summary Generation
- EG stands for Entailment Generation

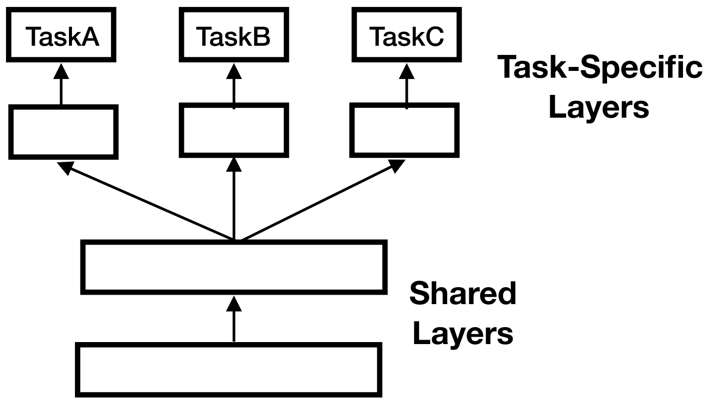
MTL Architecture (Layer Specific Sharing)



- Belinkov et al. (2017) observed that lower layers of RNN cells in a seq2seq machine translation model learn to represent word structure, while higher layers are more focused on high-level semantic meanings.
- We believe that these tasks have different training data distributions and low-level representations, they can still benefit from sharing their models' high-level components.
- Thus, we keep the lower-level layer of the 2-layer encoder/decoder of all three tasks unshared, while we share the higher layer across the three tasks.

Soft vs. Hard Parameter Sharing

- Hard-sharing: In the most common multi-task learning hard-sharing approach, the parameters to be shared are forced to be the same. As a result, gradient information from multiple tasks will directly pass through shared parameters, hence forcing a common space representation for all the related tasks.



- Soft-sharing: We encourage shared parameters to be close in representation space by penalizing their L2 distances. Unlike hard sharing, this approach gives more flexibility for the tasks by only loosely coupling the shared space representations.

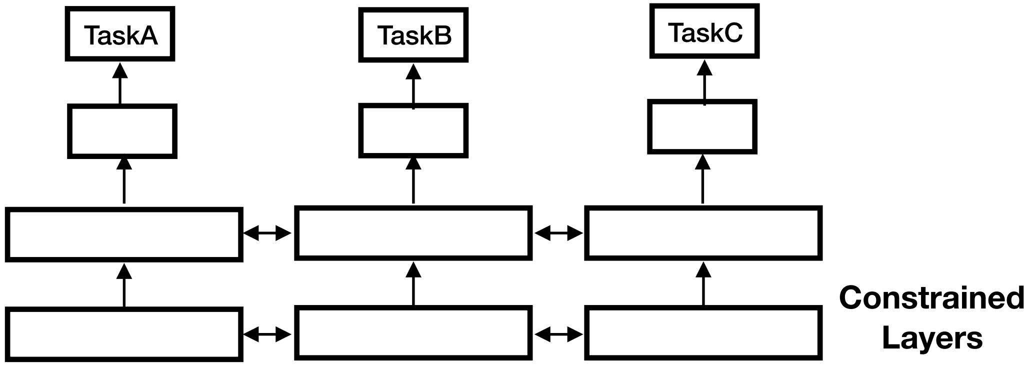


Image Credits: [7]

[Duong et al., 2015; Yang & Hospedales, 2017]

Results

Models	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
PREVIOUS WORK				
Seq2Seq(50k vocab) (See et al., 2017)	31.33	11.81	28.83	12.03
Pointer (See et al., 2017)	36.44	15.66	33.42	15.35
Pointer+Coverage (See et al., 2017) ⋆	39.53	17.28	36.38	18.72
Pointer+Coverage (See et al., 2017) †	38.82	16.81	35.71	18.14
OUR MODELS				
Two-Layer Baseline (Pointer+Coverage) ⊗	39.56	17.52	36.36	18.17
⊗ + Entailment Generation	39.84	17.63	36.54	18.61
⊗ + Question Generation	39.73	17.59	36.48	18.33
⊗ + Entailment Gen. + Question Gen.	39.81	17.64	36.54	18.54

Table: Performance of our multi-task models on CNN/DailyMail dataset (~300K examples).

* Our multi-task model is stat. signif. better than baseline (based on bootstrap test with 100K samples: Efron and Tibshirani, 1994).

Results

Models	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
PREVIOUS WORK				
Seq2Seq(50k vocab) (See et al., 2017)	31.33	11.81	28.83	12.03
Pointer (See et al., 2017)	36.44	15.66	33.42	15.35
Pointer+Coverage (See et al., 2017) ⋆	39.53	17.28	36.38	18.72
Pointer+Coverage (See et al., 2017) †	38.82	16.81	35.71	18.14
OUR MODELS				
Two-Layer Baseline (Pointer+Coverage) ⊗	39.56	17.52	36.36	18.17
⊗ + Entailment Generation	39.84	17.63	36.54	18.61
⊗ + Question Generation	39.73	17.59	36.48	18.33
⊗ + Entailment Gen. + Question Gen.	39.81	17.64	36.54	18.54

Table: Performance of our multi-task models on CNN/DailyMail dataset (~300K examples).

Human evaluation: Multi-task model is better than baseline

* Our multi-task model is stat. signif. better than baseline (based on bootstrap test with 100K samples: Efron and Tibshirani, 1994).

Results

Models	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
PREVIOUS WORK				
Seq2Seq(50k vocab) (See et al., 2017)	31.33	11.81	28.83	12.03
Pointer (See et al., 2017)	36.44	15.66	33.42	15.35
Pointer+Coverage (See et al., 2017) ⋆	39.53	17.28	36.38	18.72
Pointer+Coverage (See et al., 2017) †	38.82	16.81	35.71	18.14
OUR MODELS				
Two-Layer Baseline (Pointer+Coverage) ⊗	39.56	17.52	36.36	18.17
⊗ + Entailment Generation	39.84	17.63	36.54	18.61
⊗ + Question Generation	39.73	17.59	36.48	18.33
⊗ + Entailment Gen. + Question Gen.	39.81	17.64	36.54	18.54

Table: Performance of our multi-task models on CNN/DailyMail dataset (~300K examples).

Human evaluation: Multi-task model is better than baseline

Models	R-1	R-2	R-L
See et al. (2017)	34.30	14.25	30.82
Baseline	35.96	15.91	32.92
Multi-Task (EG + QG)	36.73	16.15	33.58

Table: Performance of various models on DUC 2002 test only setup (567 examples).

* Our multi-task model is stat. signif. better than baseline (based on bootstrap test with 100K samples: Efron and Tibshirani, 1994).

Analysis

Ground-truth: *celtic were defeated 3-2 after extra-time in the scottish cup semi-final . leigh griffiths had a goal-bound shot blocked by a clear handball. however, no action was taken against offender josh meekings . the hoops have written the sfa for an 'understanding' of the decision .*

See et al. (2017): *john hartson was once on the end of a major hampden injustice while playing for celtic . but he can not see any point in his old club writing to the scottish football association over the latest controversy at the national stadium . hartson had a goal wrongly disallowed for offside while celtic were leading 1-0 at the time but went on to lose 3-2 .*

Our Baseline: *john hartson scored the late winner in 3-2 win against celtic . celtic were leading 1-0 at the time but went on to lose 3-2 . some fans have questioned how referee steven mclean and additional assistant alan muir could have missed the infringement .*

Multi-task: *celtic have written to the scottish football association in order to gain an ' understanding ' of the refereeing decisions . the hoops were left outraged by referee steven mclean 's failure to award a penalty or red card for a clear handball in the box by josh meekings . celtic striker leigh griffiths has a goal-bound shot blocked by the outstretched arm of josh meekings .*

* Our multi-task models are stat. signif. better than baselines.

Reinforcement Learning

Reinforcement Learning

- Reinforcement Learning (RL) is a training mechanism in which an agent or a policy is allowed to interact with a given environment in order to maximize a reward.

- RL has successful application to many research areas such as continuous control, dialogue systems, and games.

- Recently, a special case of RL called policy gradients based reinforcement learning, has been widely applied to text generation problems in NLP through REINFORCE algorithm.

[Williams, 1992; Tesauro, 1995; White & Sofge, 1992; Singh et al., 2002; Ren et al., 2017; Rennie et al., 2017; Paulus et al., 2018; Chen & Bansal, 2018; Celikyilmaz et al., 2018]

REINFORCE

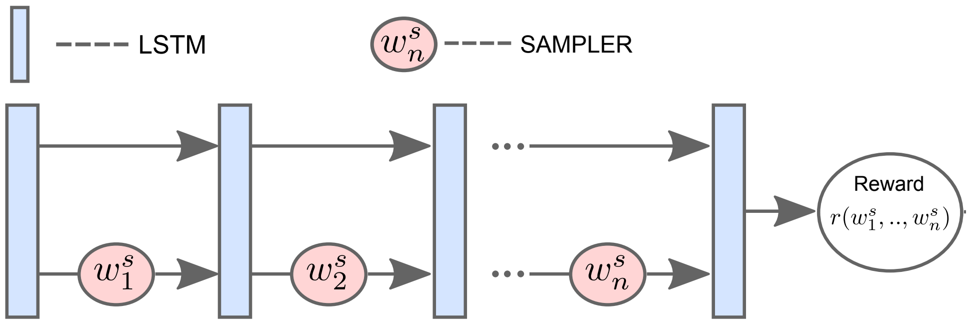


Figure: Overview of an LSTM decoder with sampling of words in a sequential fashion to generate a sentence. We measure a reward for the generated sentence w.r.t. the ground-truth and use this reward to update RL policy (model).

[Williams, 1992]

REINFORCE

Loss function:

$$L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)]$$

Gradient estimation:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}\left[r(w^s)\nabla_\theta \log p_\theta(w^s)\right].$$

Gradient approximation:

$$\nabla_\theta L(\theta) \approx -r(w^s)\nabla_\theta \log p_\theta(w^s).$$

Reducing variance

$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}\left[(r(w^s) - b)\nabla_\theta \log p_\theta(w^s)\right].$$

Mixed loss:

$$L_{\text{MIXED}} = (1-\gamma)L_{\text{XE}} + \gamma L_{\text{RL}}$$

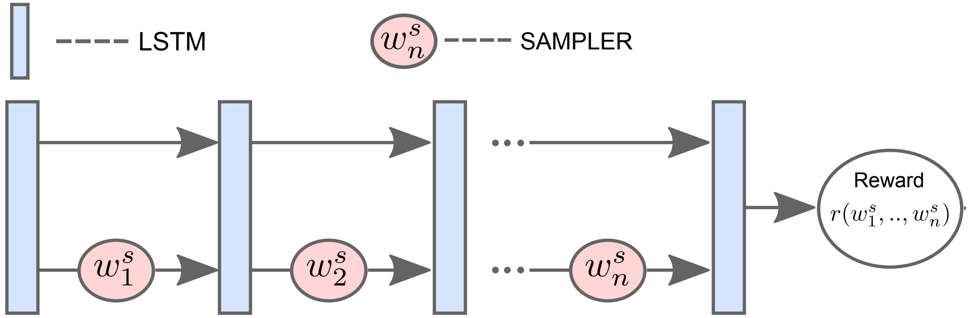


Figure: Overview of an LSTM decoder with sampling of words in a sequential fashion to generate a sentence. We measure a reward for the generated sentence w.r.t. the ground-truth and use this reward to update RL policy (model).

[Williams, 1992]

RL for Abstractive Summarization

We address *three* important aspects (saliency, directed logical entailment/correctness, and non-redundancy) of a good abstractive text summary via reinforcement learning approach with *two novel reward functions*.

[See et al., 2017; Paulus et al., 2017; Celikyilmaz et al., 2018]

RL for Abstractive Summarization

We address *three* important aspects (saliency, directed logical entailment/correctness, and non-redundancy) of a good abstractive text summary via reinforcement learning approach with *two novel reward functions*.

We also introduce a *novel and effective multi-reward approach* of optimizing multiple rewards simultaneously in alternate multi-task mini-batches.

[See et al., 2017; Paulus et al., 2017; Celikyilmaz et al., 2018]

Reward Functions

Rouge Reward

Based on the primary summarization metric of ROUGE package (Lin, 2004).

Reward Functions

Rouge Reward

Based on the primary summarization metric of ROUGE package (Lin, 2004).

Saliency Reward

Gives *higher* weight to the important, salient words/phrases when calculating the ROUGE score.

Entailment Reward

Based on whether each sentence of the generated summary is entailed by the ground-truth summary.

Saliency Reward: ROUGESal

ROUGESal reward gives higher weight to the important, salient words/phrases when calculating the ROUGE score (which by default assumes all words are equally weighted):

- To learn these saliency weights, we train our saliency predictor on {sentence, answer spans} pairs from the popular SQuAD reading comprehension dataset (Rajpurkar et al., 2016) (Wiki domain).
- We treat the human-annotated answer spans for important questions as representative salient information in the document.
- This saliency predictor is run on the ground-truth summary to get an importance weight for each word (used in ROUGE matching).

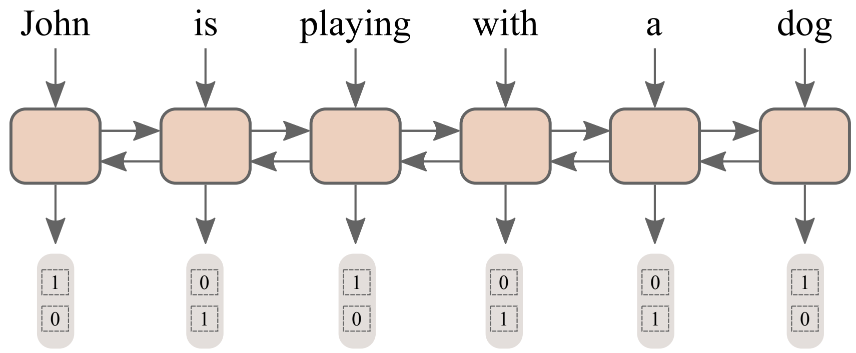


Figure: Overview of our saliency prediction model.

Entailment Reward: Entail

- A good summary should be logically entailed by source document, i.e., have no contradictory/unrelated information. We use an entailment scorer and its multi-sentence, length-normalized extension (to avoid very short sentences achieving misleadingly high entailment scores) as our "Entail" reward.
- We train the entailment classifier (Parikh et al., 2016) on the SNLI and Multi-NLI datasets and calculate the entailment probability score between the ground-truth (GT) summary (as premise) and each sentence of the generated summary (as hypothesis), and use average score as our Entail reward.

$$\text{Entail} = \text{Entail} \times \frac{\#\text{tokens in generated summary}}{\#\text{tokens in reference summary}}$$

Multi-Reward Optimization

- One approach for multi-reward optimization is to use a weighted combination of the rewards, but this has the issue of finding the complex scaling and weight balance among these diverse reward combinations.

Multi-Reward Optimization

- One approach for multi-reward optimization is to use a weighted combination of the rewards, but this has the issue of finding the complex scaling and weight balance among these diverse reward combinations.

- To address this issue, we instead introduce a simple multi-reward optimization approach inspired from multi-task learning, where we have different tasks, and they share all model parameters while having their own optimization function (different reward functions in this case), with alternate mini-batches:

$$L_{\mathrm{RL}_1} = -(r_1(w^s) - r_1(w^a))\nabla_\theta \log p_\theta(w^s)$$
$$L_{\mathrm{RL}_2} = -(r_2(w^s) - r_2(w^a))\nabla_\theta \log p_\theta(w^s)$$

Results (CNN/Daily Mail)

Models	R-1	R-2	R-L	M
PREVIOUS WORK				
Nallapati (2016)⋆	35.46	13.30	32.65	-
See et al. (2017)	39.53	17.28	36.38	18.72
Paulus (2017) (XE)⋆	38.30	14.81	35.49	-
Paulus (2017) (RL)⋆	39.87	15.82	36.90	-
OUR MODELS				
Baseline (XE)	39.41	17.33	36.07	18.27
ROUGE (RL)	39.99	17.72	36.66	18.93
Entail (RL)	39.53	17.51	36.44	20.15
ROUGESal (RL)	40.36	17.97	37.00	19.84
ROUGE+Ent (RL)	40.37	17.89	37.13	19.94
ROUGESal+Ent (RL)	40.43	18.00	37.10	20.02

Table: Results on CNN/Daily Mail (nonanonymous). * represents previous work on anonymous version. 'XE': cross-entropy loss, 'RL': reinforce mixed loss (XE+RL). Columns 'R': ROUGE, 'M': METEOR. Final multi-reward RL model improvements are statistically significant over baseline, ROUGE-RL, Entail-RL.

Results (CNN/Daily Mail)

Models	R-1	R-2	R-L	M
PREVIOUS WORK				
Nallapati (2016)*	35.46	13.30	32.65	-
See et al. (2017)	39.53	17.28	36.38	18.72
Paulus (2017) (XE)*	38.30	14.81	35.49	-
Paulus (2017) (RL)*	39.87	15.82	36.90	-
OUR MODELS				
Baseline (XE)	39.41	17.33	36.07	18.27
ROUGE (RL)	39.99	17.72	36.66	18.93
Entail (RL)	39.53	17.51	36.44	20.15
ROUGESal (RL)	40.36	17.97	37.00	19.84
ROUGE+Ent (RL)	40.37	17.89	37.13	19.94
ROUGESal+Ent (RL)	40.43	18.00	37.10	20.02

Table: Results on CNN/Daily Mail (nonanonymous). * represents previous work on anonymous version. 'XE': cross-entropy loss, 'RL': reinforce mixed loss (XE+RL). Columns 'R': ROUGE, 'M': METEOR. Final multi-reward RL model improvements are statistically significant over baseline, ROUGE-RL, Entail-RL.

Results (CNN/Daily Mail)

Models	R-1	R-2	R-L	M
PREVIOUS WORK				
Nallapati (2016)*	35.46	13.30	32.65	-
See et al. (2017)	39.53	17.28	36.38	18.72
Paulus (2017) (XE)*	38.30	14.81	35.49	-
Paulus (2017) (RL)*	39.87	15.82	36.90	-
OUR MODELS				
Baseline (XE)	39.41	17.33	36.07	18.27
ROUGE (RL)	39.99	17.72	36.66	18.93
Entail (RL)	39.53	17.51	36.44	20.15
ROUGESal (RL)	40.36	17.97	37.00	19.84
ROUGE+Ent (RL)	40.37	17.89	37.13	19.94
ROUGESal+Ent (RL)	40.43	18.00	37.10	20.02

Human evaluation:
Our Multi-reward model is better than baseline

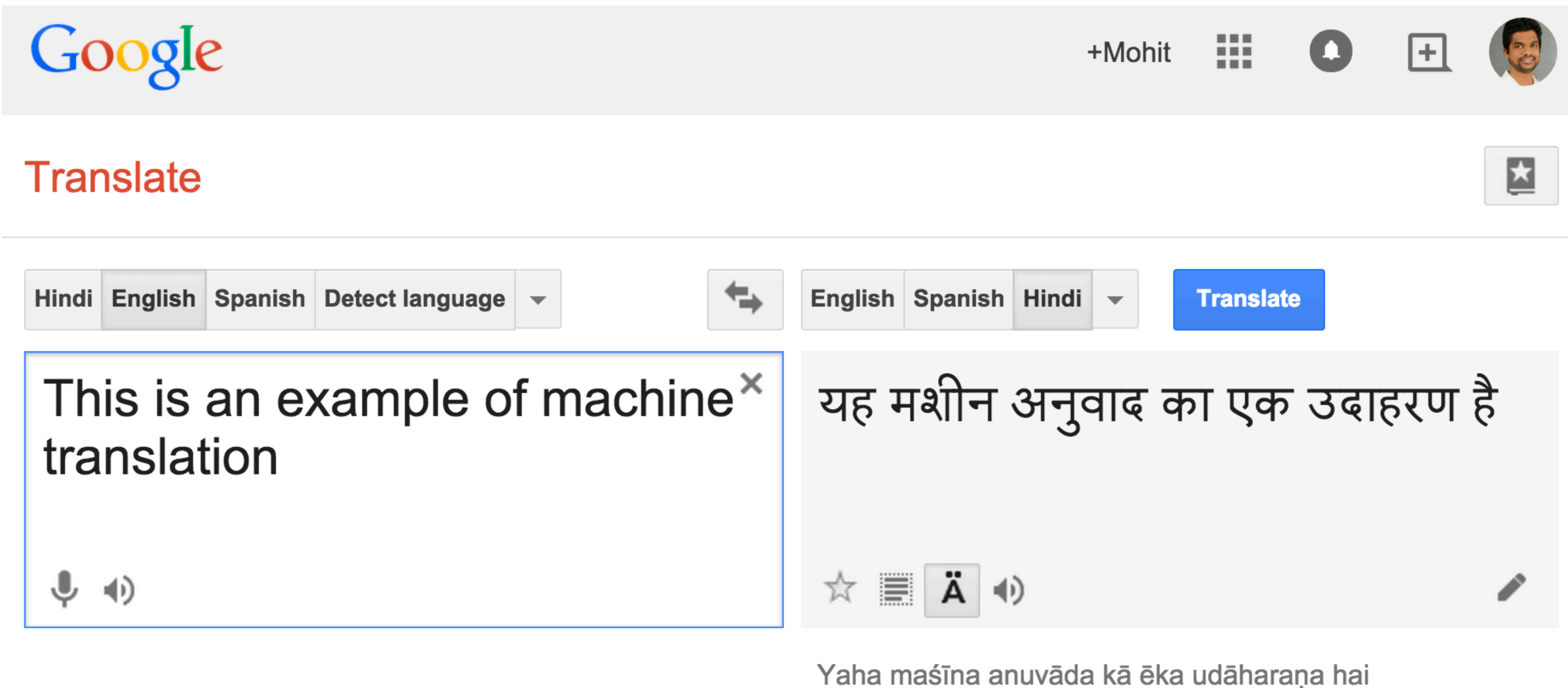
Table: Results on CNN/Daily Mail (nonanonymous). * represents previous work on anonymous version. 'XE': cross-entropy loss, 'RL': reinforce mixed loss (XE+RL). Columns 'R': ROUGE, 'M': METEOR. Final multi-reward RL model improvements are statistically significant over baseline, ROUGE-RL, Entail-RL.

Thank You

Machine Translation

Machine Translation

- Useful for tons of companies, online traffic, and our international communication!



Statistical Machine Translation

- Source language f (e.g., French)
- Target language e (e.g., English)
- We want the best target (English) translation given the source (French) input sentence, hence the probabilistic formulation is:

$$\hat{e} = \mathrm{argmax}_e p(e|f)$$

- Using Bayes rule, we get the following (since p(f) in the denominator is independent of the argmax over e):

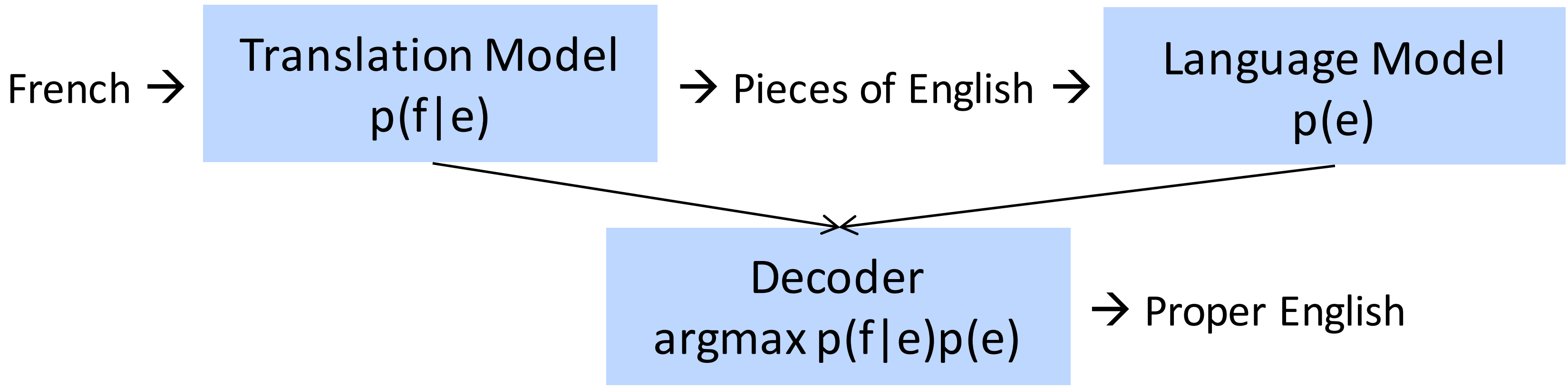
$$\hat{e} = \mathrm{argmax}_e p(e|f) = \mathrm{argmax}_e p(f|e)p(e)$$

[Richard Socher CS224d]

Statistical Machine Translation

- The first part is known as the 'Translation Model' p(f|e) and is trained on parallel corpora of {f,e} sentence pairs, e.g., from EuroParl or Canadian parliament proceedings in multiple languages

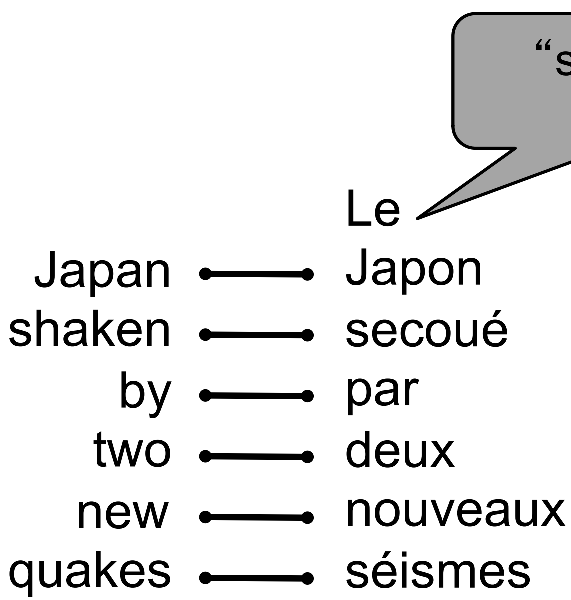
- The second part p(e) is the 'Language Model' and can be trained on tons more monolingual data, which is much easier to find!

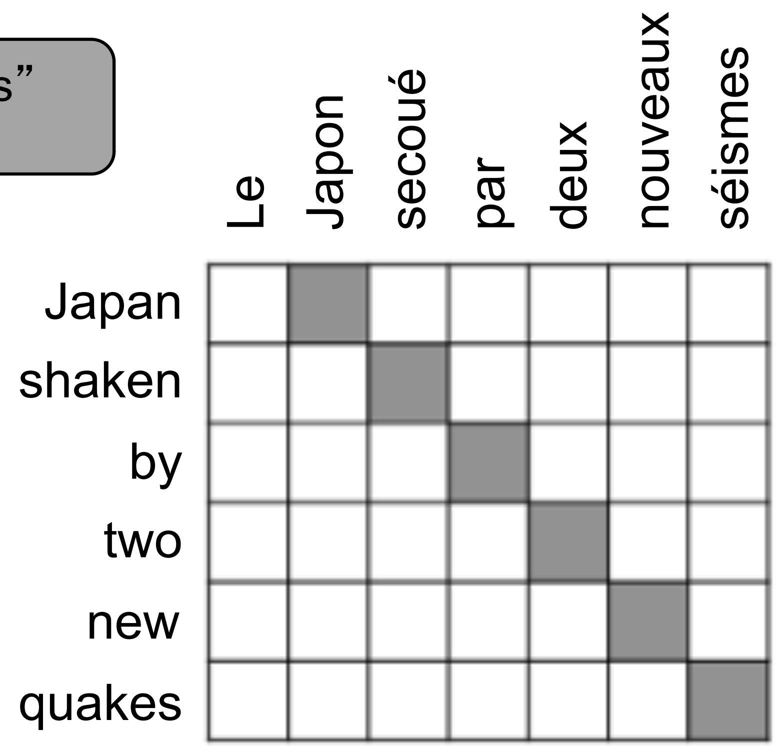


[Richard Socher CS224d]

Statistical Machine Translation

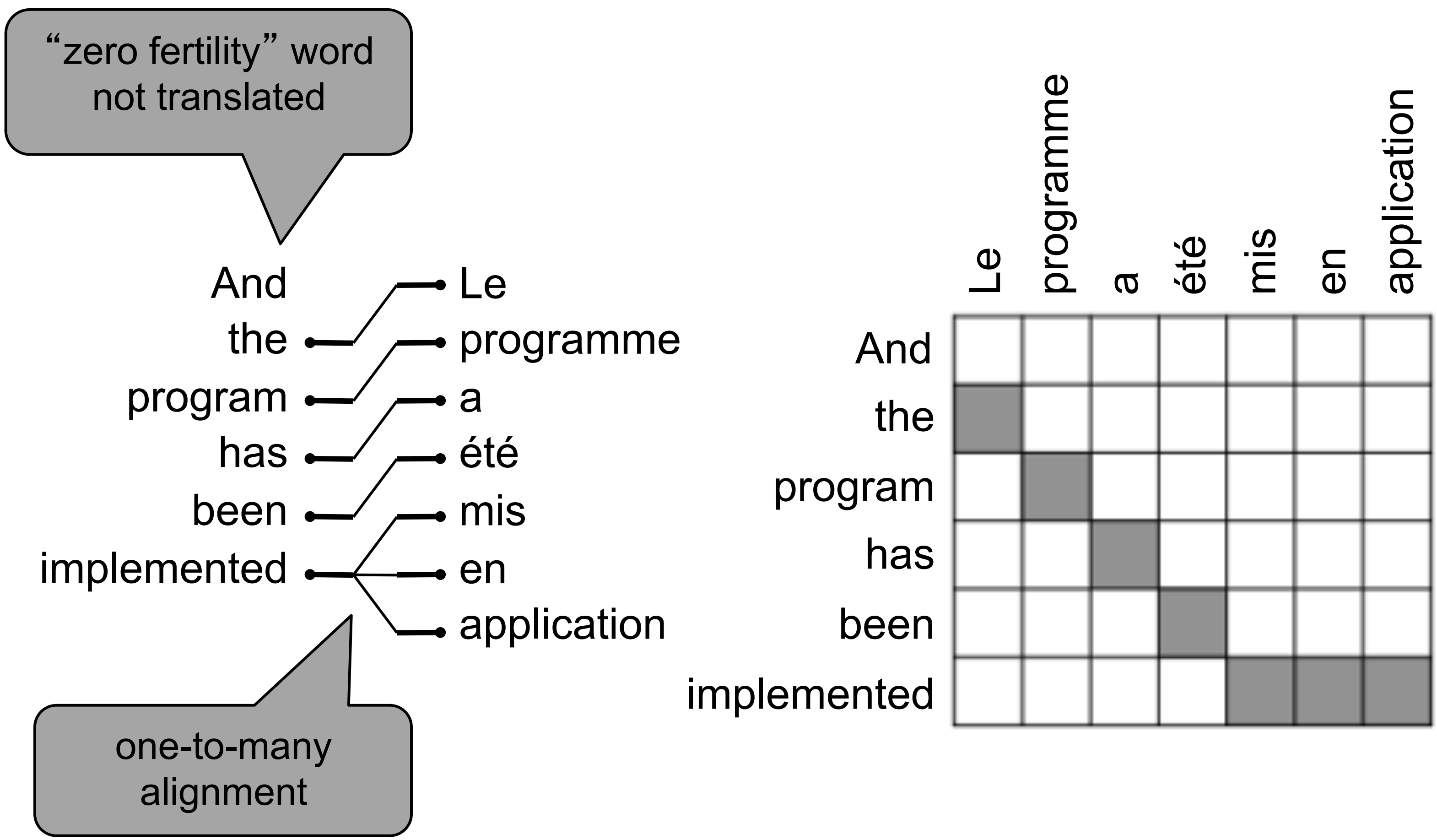
- First step in traditional machine translation is to find alignments or translational matchings between the two sentences, i.e., predict which words/phrases in French align to which words/phrases in English.
- Challenging problem: e.g., some words may not have any alignments:





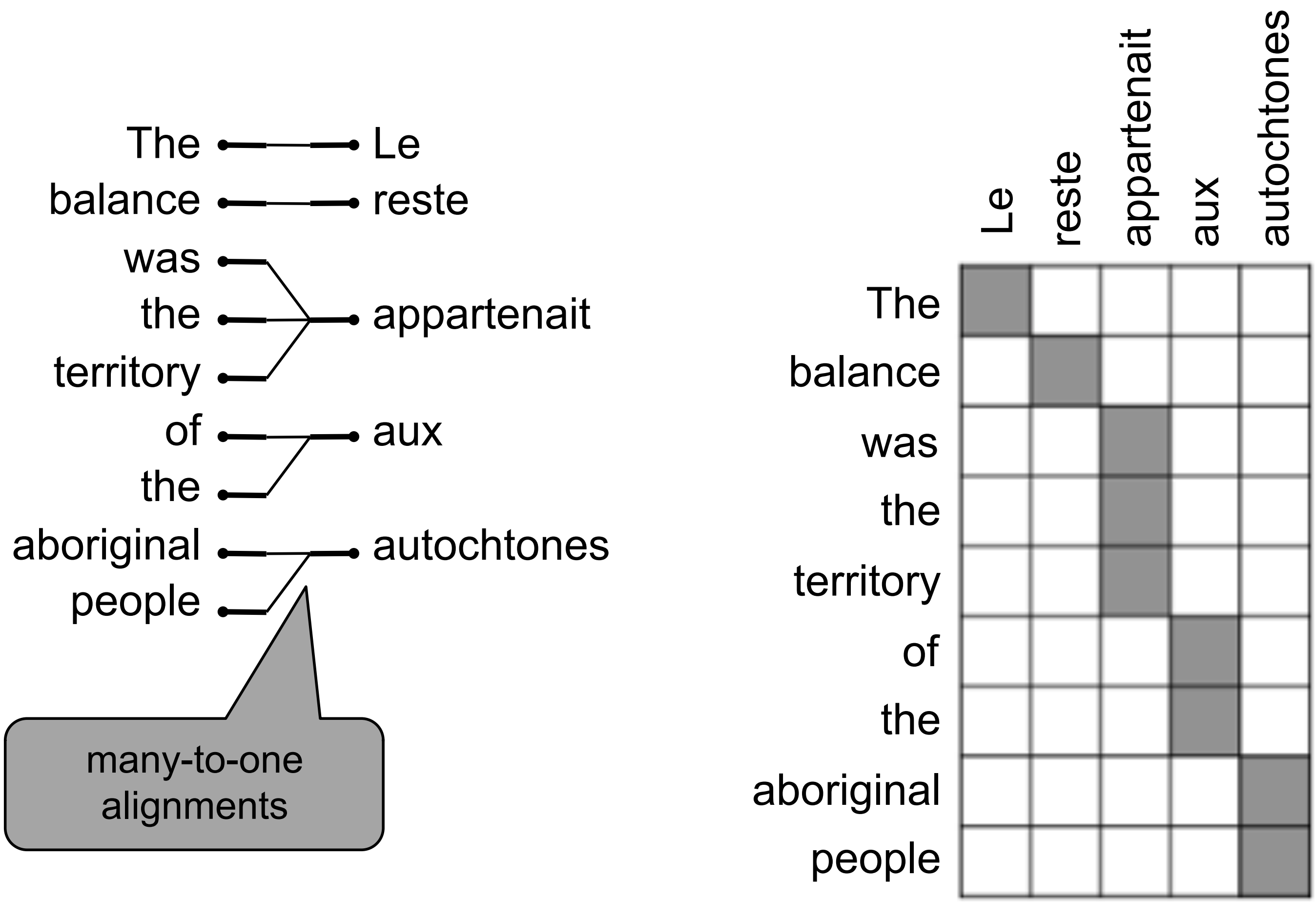
Statistical Machine Translation

- One word in the source sentence might align to several words in the target sentence:



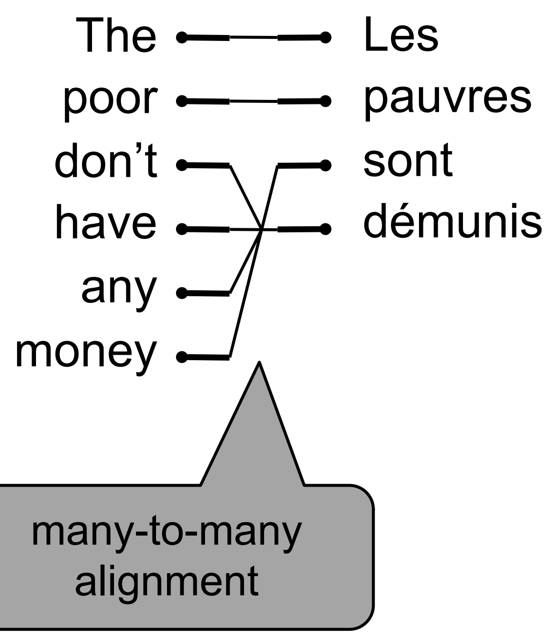
Statistical Machine Translation

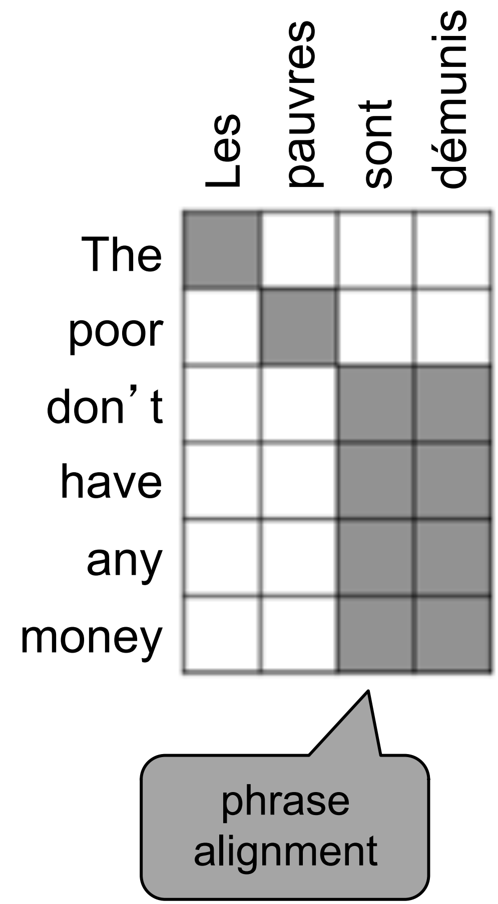
- Many words in the source sentence might align to a single word in the target sentence:



Statistical Machine Translation

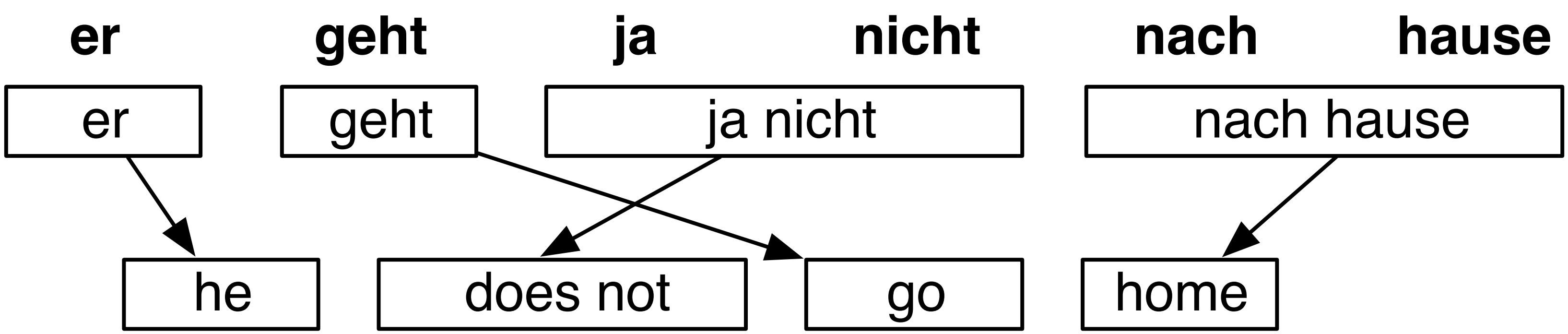
- And finally, many words in the source sentence might align to many words in the target sentence:





Statistical Machine Translation

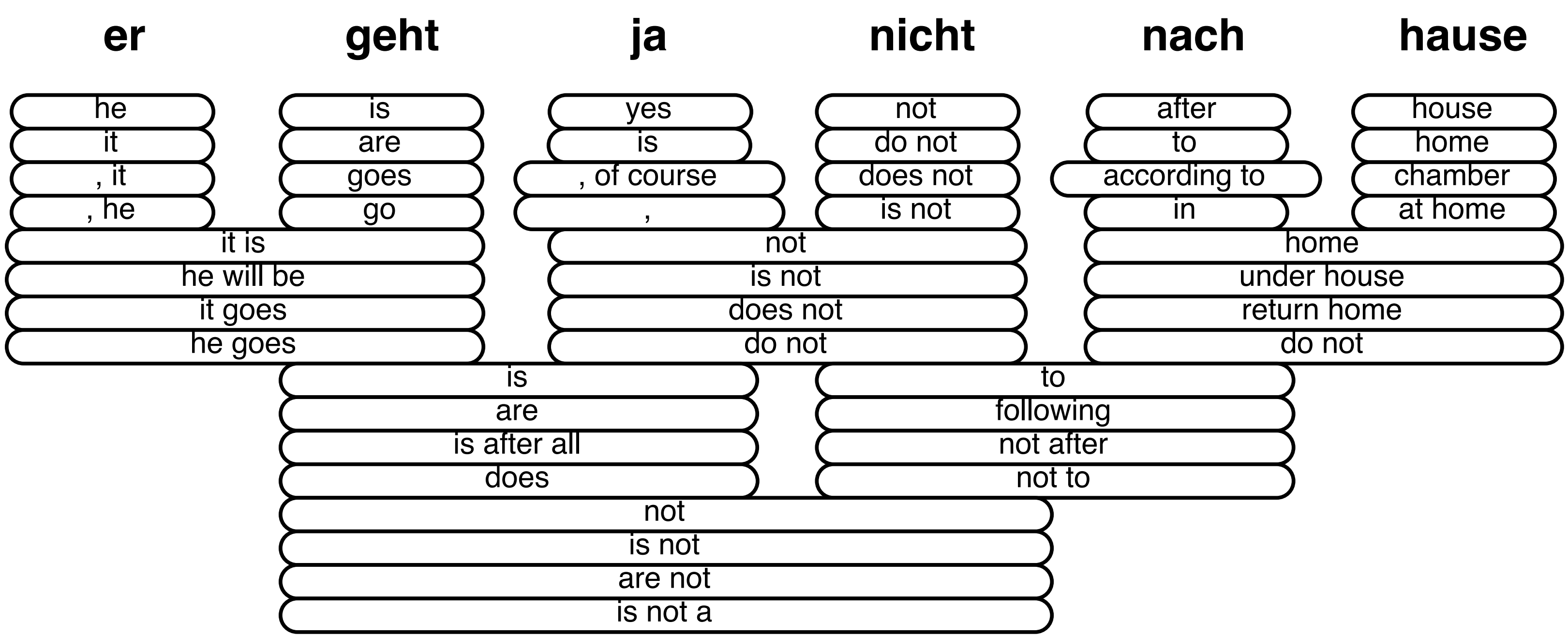
- After learning the word and phrase alignments, the model also needs to figure out the reordering, esp. important in language pairs with very different orders!



Statistical Machine Translation

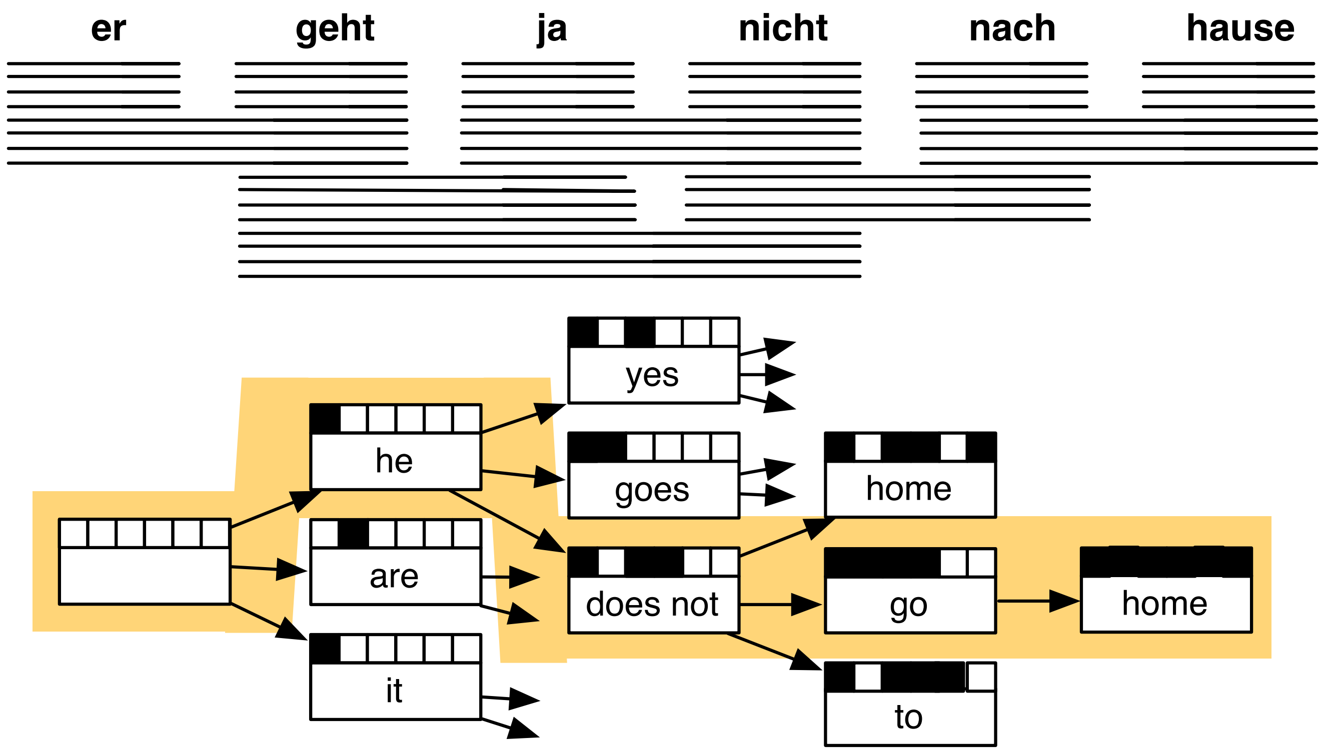
- After many steps, you get the large ‘phrase table’. Each phrase in the source language can have many possible translations in the target language, and hence the search space can be combinatorially large!

Translation Options

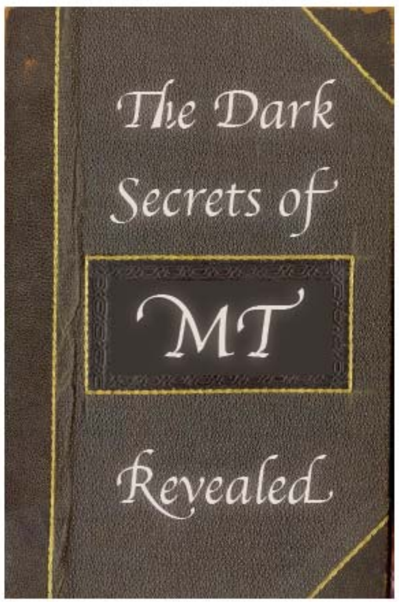


Statistical Machine Translation

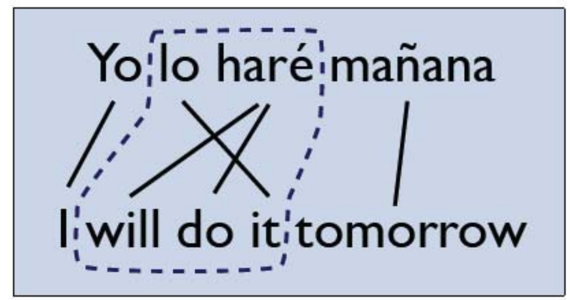
- Finally, you decode this hard search problem to find the best translation, e.g., using beam search on the several combinatorial paths through this phrase table (and also include the language model p(e) to rerank)



Alignment Model Details



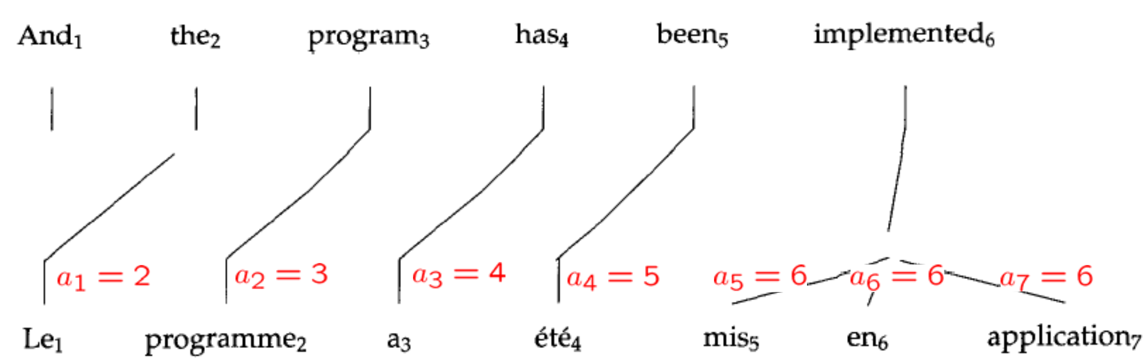
1. *Align words with a probabilistic model*
2. *Infer presence of larger structures from this alignment*
3. *Translate with the larger structures*



IBM Model 1

- Alignments: a hidden vector called an alignment specifies which English source is responsible for each French target word.
- The first, simplest IBM model treated alignment probabilities as roughly uniform:

$$a = a_1 \dots a_J$$



$$P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i)$$

$$= \prod_j \frac{1}{I+1} P(f_j|e_i)$$

$$P(f|e) = \sum_a P(f, a|e)$$

[Brown et al., 1993]

IBM Model 2 (Distortion)

- The next more advanced model captures the notion of ‘distortion’, i.e., how far from the diagonal is the alignment

$$P(f, a|e) = \prod_j P(a_j = i|j, I, J)P(f_j|e_i)$$

$$P(dist = i - j\frac{I}{J})$$

$$\frac{1}{Z}e^{-\alpha(i - j\frac{I}{J})}$$

- Other approaches for biasing alignment towards diagonal include relative vs absolute alignment, asymmetric distances, and learning a full multinomial over distances

[Brown et al., 1993]

IBM Models 1/2 EM Training

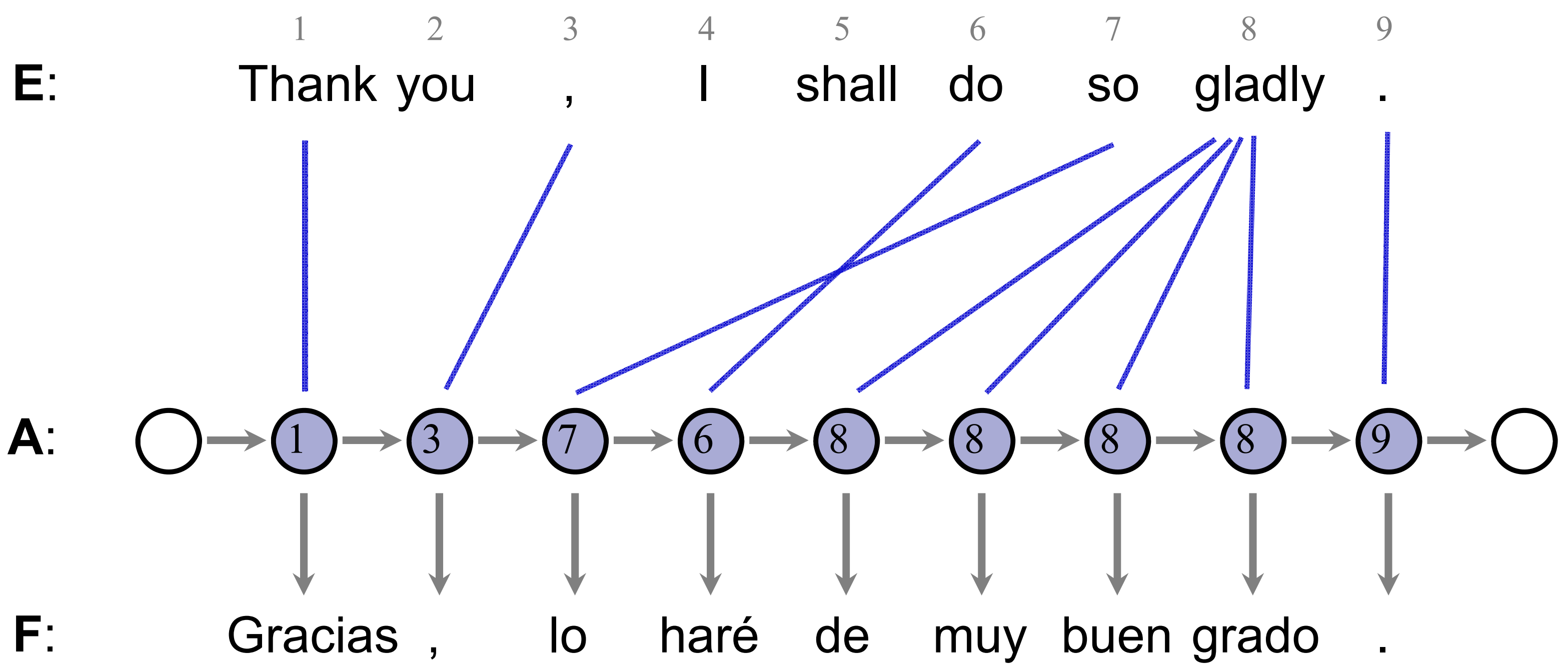
- Model Parameters:
  - Translational Probabilities: $P(f_j|e_i)$
  - Distortion Probabilities: $P(a_j = i|j, I, J)$
- Start with uniform P(f_j | e_i) parameters, including P(f_j | null)
- For each sentence in training corpus:
  - For each French position j:
    - Calculate posterior over English positions using:

$$P(a_j = i|f, e) = \frac{P(a_j = i|j, I, J)P(f_j|e_i)}{\sum_{i'} P(a_j = i'|j, I, J)P(f_j|e'_i)}$$

    - Increment count of word f_j with word e_i by these amounts
    - Similarly re-estimate distortion probabilities for Model2
- Iterate until convergence

[Brown et al., 1993]

HMM Model

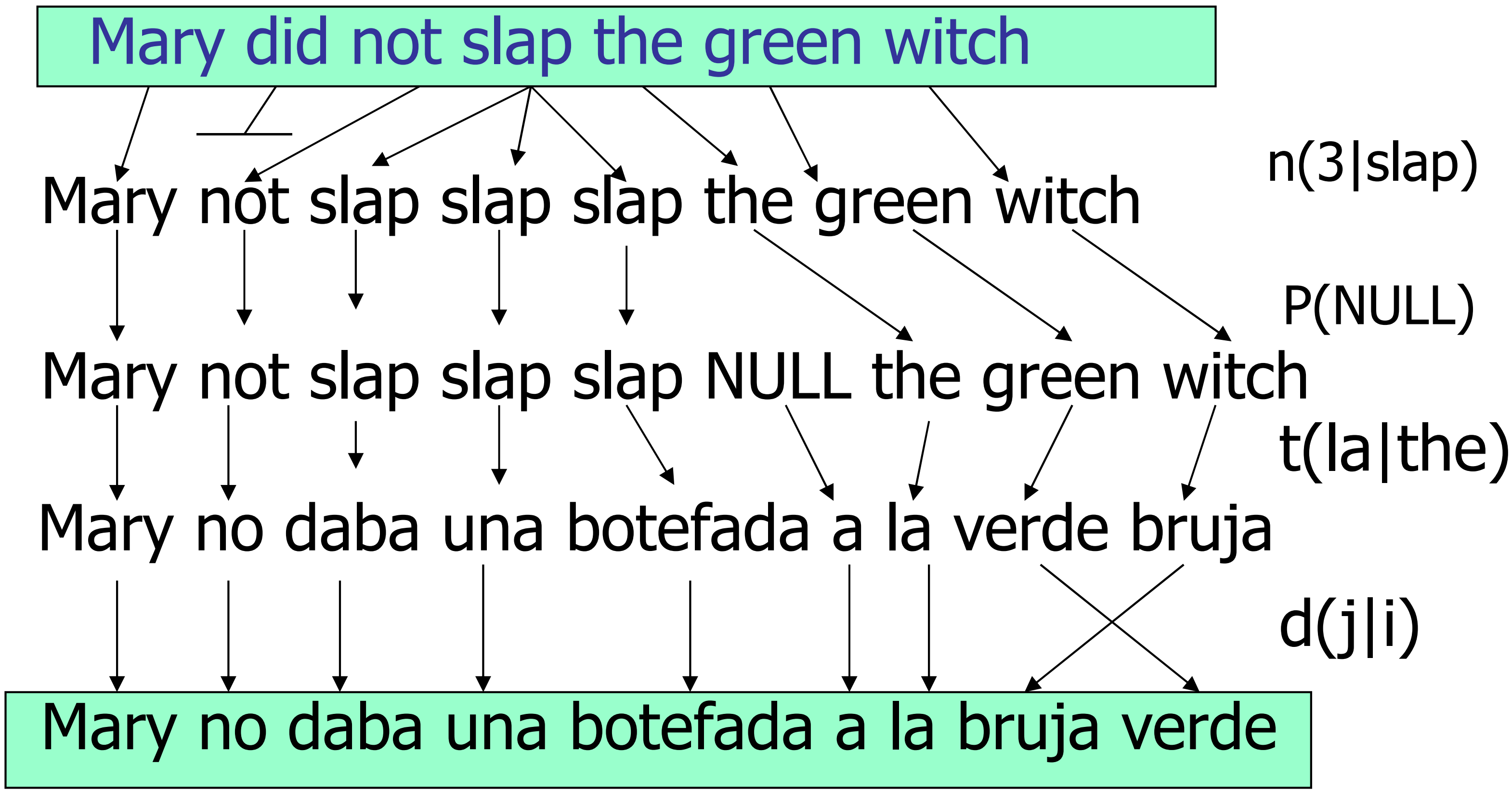


Model Parameters

Emissions: P(F_1 = Gracias | E_{A_1} = Thank) *Transitions*: P(A_2 = 3 | A_1 = 1)

[Vogel et al., 1996]

IBM Models 3/4/5 (Fertility)



[Vogel et al., 1996]

IBM Models 3/4/5 (Fertility)

the

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
le	0.497	1	0.746
la	0.207	0	0.254
les	0.155		
l'	0.086		
ce	0.018		
cette	0.011		

not

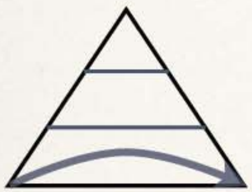
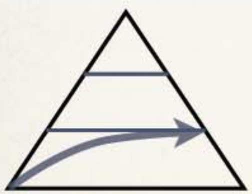
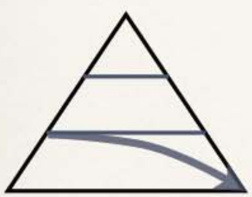
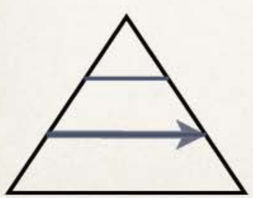
f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
ne	0.497	2	0.735
pas	0.442	0	0.154
non	0.029	1	0.107
rien	0.011		

farmers

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
agriculteurs	0.442	2	0.731
les	0.418	1	0.228
cultivateurs	0.046	0	0.039
producteurs	0.021		

[Vogel et al., 1996]

Syntactic Machine Translation

	string-to-string	ITG (Wu 1997)	Hiero (Chiang 2005)
	string-to-tree	Yamada & Knight 2001	Galley et al 2004/2006
	tree-to-string		Huang et al 2006 Y Liu et al 2006
	tree-to-tree	DOT (Poutsma 2000) Eisner 2003	Stat-XFER (Lavie et al 2008) M Zhang et al. 2008 Y Liu et al., 2009

Hiero

$$S \rightarrow \langle S_{\boxed{1}} X_{\boxed{2}}, S_{\boxed{1}} X_{\boxed{2}} \rangle$$

$$S \rightarrow \langle X_{\boxed{1}}, X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle \text{yu } X_{\boxed{1}} \text{ you } X_{\boxed{2}}, \text{have } X_{\boxed{2}} \text{ with } X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle X_{\boxed{1}} \text{ de } X_{\boxed{2}}, \text{the } X_{\boxed{2}} \text{ that } X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle X_{\boxed{1}} \text{ zhiyi, one of } X_{\boxed{1}} \rangle$$

$$X \rightarrow \langle \text{Aozhou, Australia} \rangle$$

$$X \rightarrow \langle \text{shi, is} \rangle$$

$$X \rightarrow \langle \text{shaoshu guojia, few countries} \rangle$$

$$X \rightarrow \langle \text{bangjiao, diplomatic relations} \rangle$$

$$X \rightarrow \langle \text{Bei Han, North Korea} \rangle$$

From [Chiang et al, 2005]

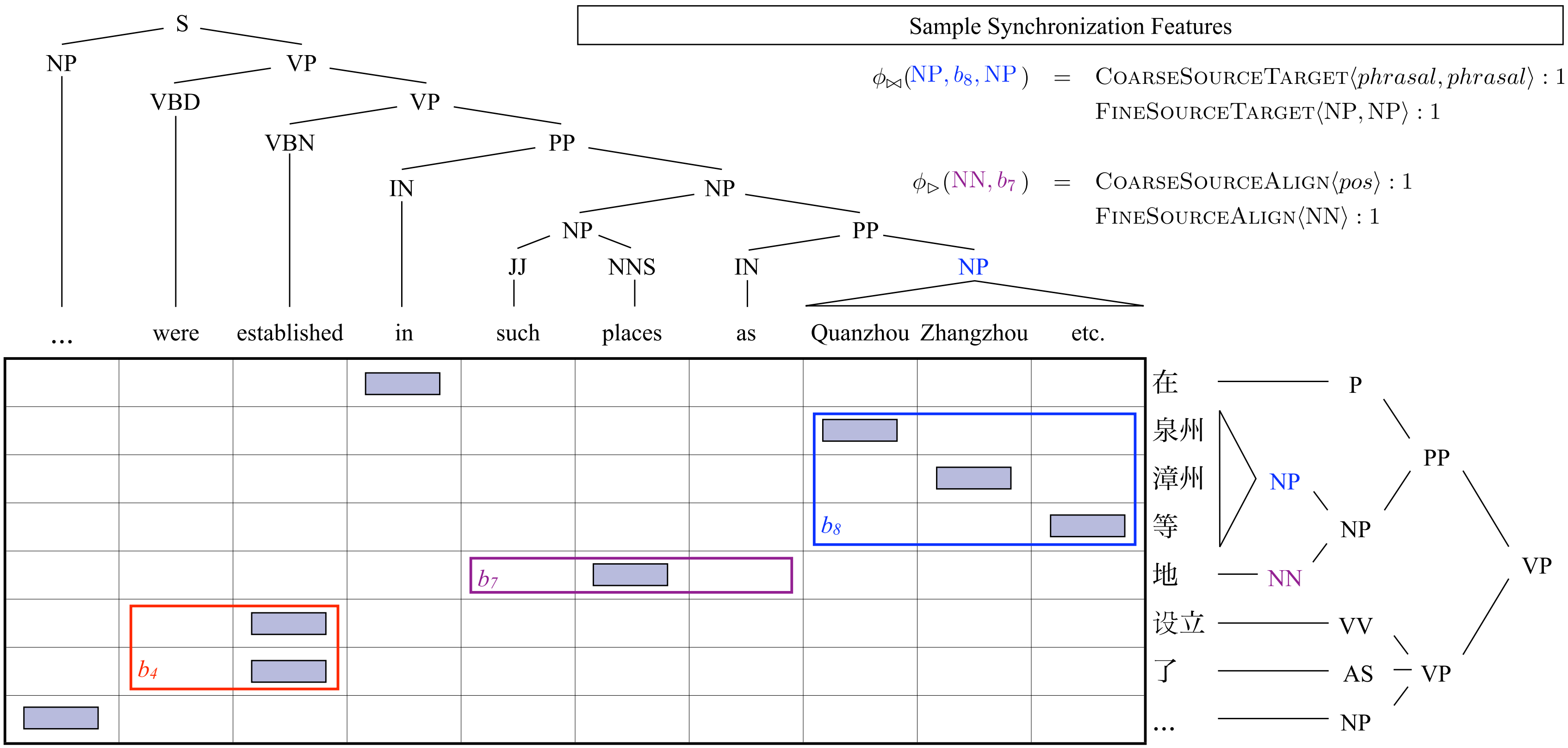
Synchronous Tree-Substitution Grammars

STSG extraction

1. Phrases
   - respect word alignments
   - are syntactic constituents on *both* sides
2. Phrase pairs form rules
3. Subtract phrases to form rules

[Shieber, 2004; Graehl et al., 2008]

Joint Parsing and Alignment



[Burkett et al., 2011]

Neural Machine Translation (next week)