Building linguistically informed models for low-resource settings

Nanyun (Violet) Peng

Collaborations with Hoifung Poon, Chris Quirk, Kristina Toutanova, Wen-tau Yih, Max Ma, Eduard Hovy, Wasi Uddin Ahmad, Zhisong Zhang, Kai-Wei Chang, Ralph Weischedel, Kevin Knight, Lili Yao, Rui Yan

Microsoft Research · Carnegie Mellon University Language Technologies Institute · UCLA · Peking University 1898

Low Resource Languages

- 25 اکتوبر، سہالے ورک زیوڈے ایتھوپیا کی صدر منتخب ہوئیں۔ سہالے ورک براعظم افریقا کی کسی بھی ریاست

- অ্যাঞ্জেলিনা জোলি একজন জনপ্রিয় মার্কিন চলচ্চিত্র অভিনেত্রী। যুক্তরাষ্ট্রের ক্যালিফোর্নিয়ার লস অ্যাঞ্জেলেসের একটি সংস্কৃতিমনা পরিবারে এই অস্কারজয়ী

Low Resource Domains

T790M is present as a minor clone in NSCLC ,

and may be selected for during therapy .

This mutation has been shown to prevent the activation

of BIM in response to gefitinib

but can be overcome by an irreversible inhibitor of EGFR.

Low Resource Domains

present

T790M is present as a minor clone in NSCLC ,

and may be selected for during therapy .

This mutation has been shown to prevent the activation

respond

of BIM in response to gefitinib

but can be overcome by an irreversible inhibitor of EGFR.

Low Resource Tasks

- Creative Composition
  - Poetry

    Two roads diverged in a yellow wood,

    And sorry I could not travel both

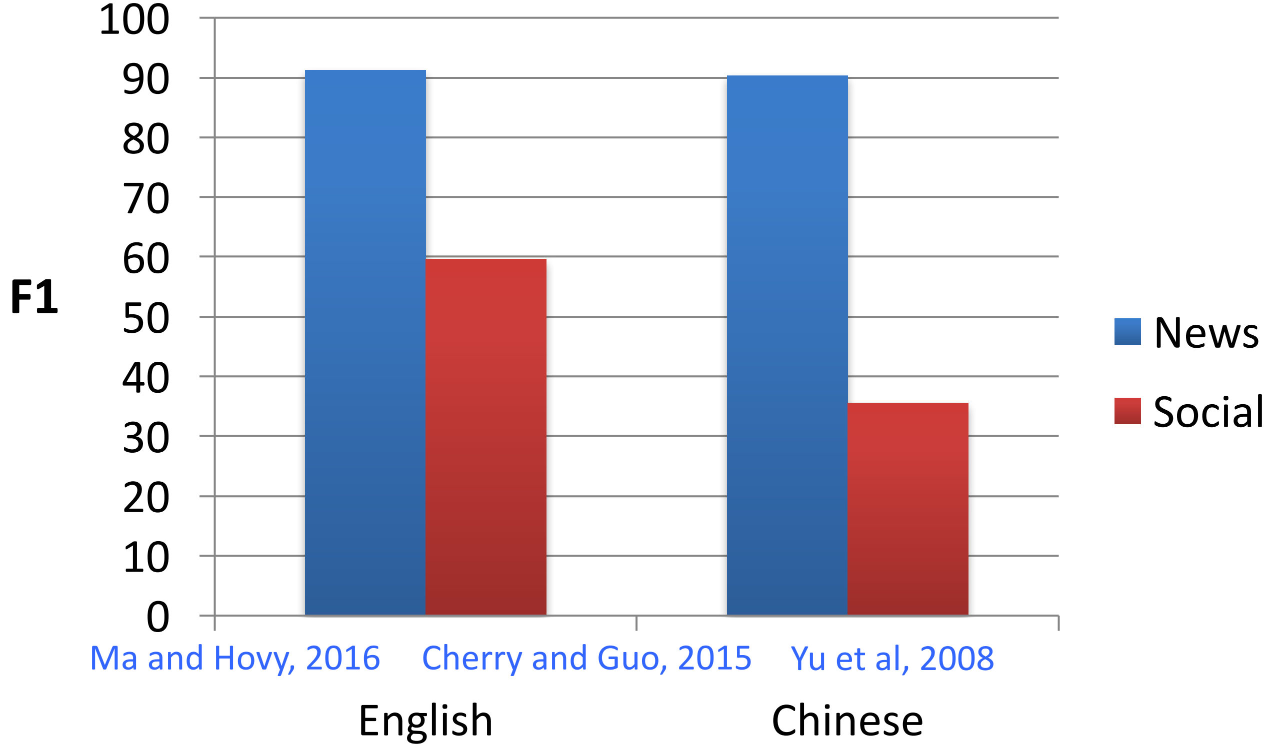
    And be one traveler, long I stood ……
  - Pun

    The magician got so mad he pulled his hare out.
  - Story

    The last person on Earth was alone in a room. There was a knock on the door...

Challenges in Low-Resource Settings

- HUGE gap on social media (low-resource) v.s news (high-resource) text:
    - informal language and insufficient annotations.



Challenges of Obtaining Training Data

- Constructing data sets is labor intensive
- Many different
  - Languages
  - Domains
  - Tasks
  - …

Building Robust Models For Low-Resource Settings

- Cross-Sentence N-ary Relation Extraction for Biomedical Domain (low resource domain)

- On Difficulties of Cross-lingual Transfer (low resource languages)

- Plan-and-Write Story Generation (low resource task)

Cross-Sentence N-ary Relation Extraction

PubMed

Mutation

T790M is present as a minor clone in NSCLC ,

and may be selected for during therapy .

This mutation has been shown to prevent the

Drug

activation of BIM in response to gefitinib but can

Gene

be overcome by an irreversible inhibitor of EGFR.

Peng et. al. (TACL2017)

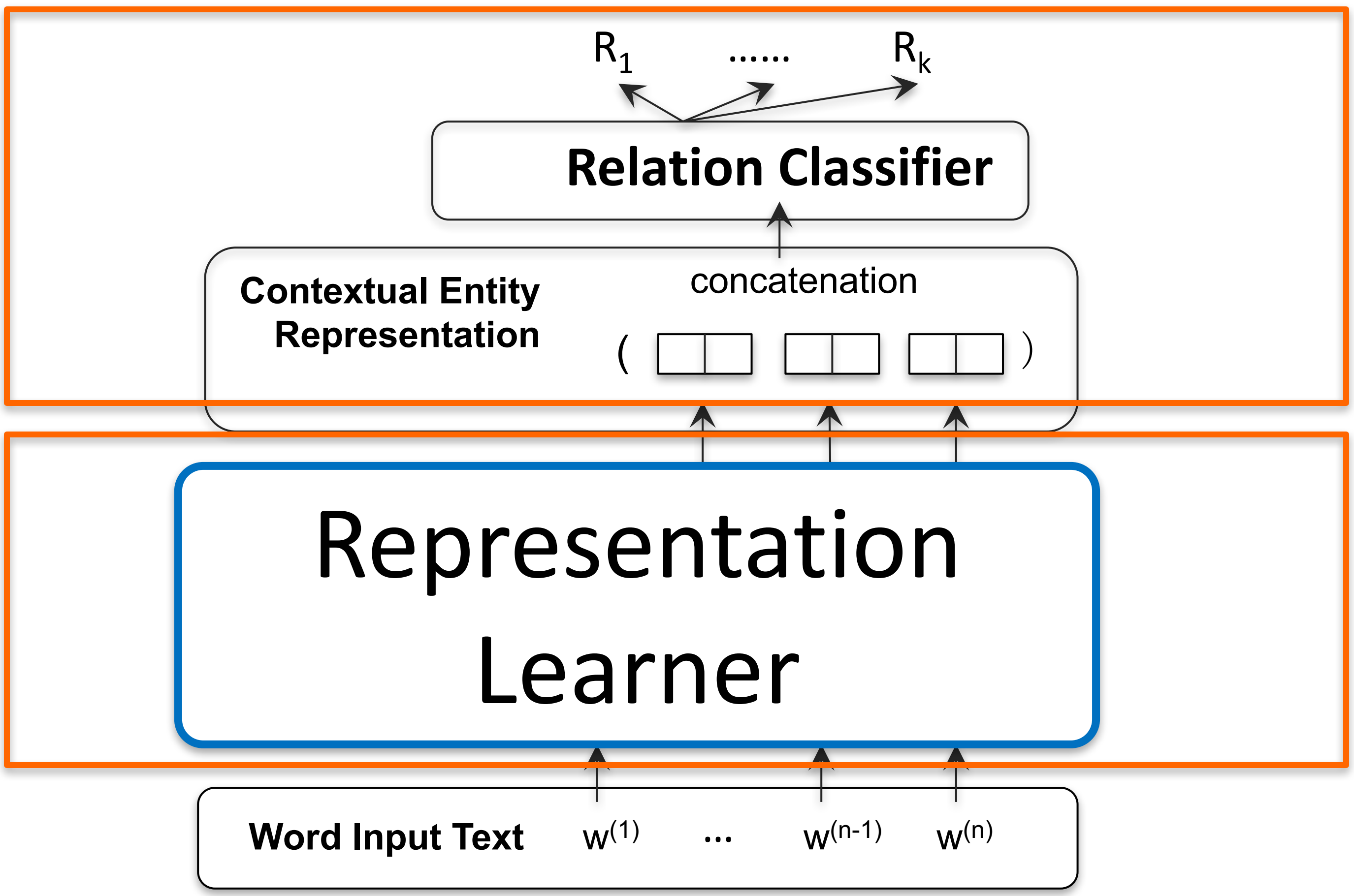
Knowledge Bases for Drug-Gene-Mutation Interaction

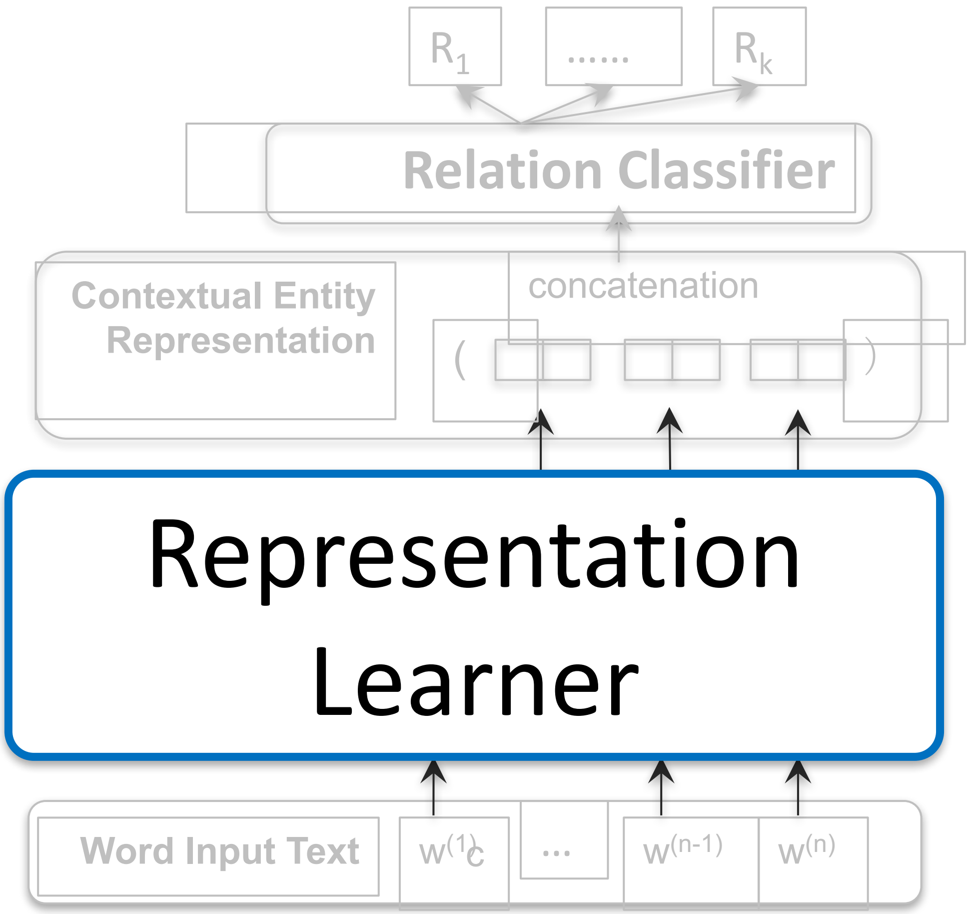
- People manually curate drug-gene-mutation interaction databases for precision medicine:
  - Gene Drug Knowledge Database (GDKD) (Dienstmann et al., 2015)
  - Clinical Interpretations of Variants in Cancer (CiViC) (Washington University School of Medicine)

Special Challenges

- N-ary relations:
  - Traditional feature-based classification method usually use features defined on the *shortest syntactic dependency paths* between two entities.
  - Such features are hard to define in the *N*-ary case.
- Cross sentence relations:
  - Traditional features become sparser and learning becomes harder.

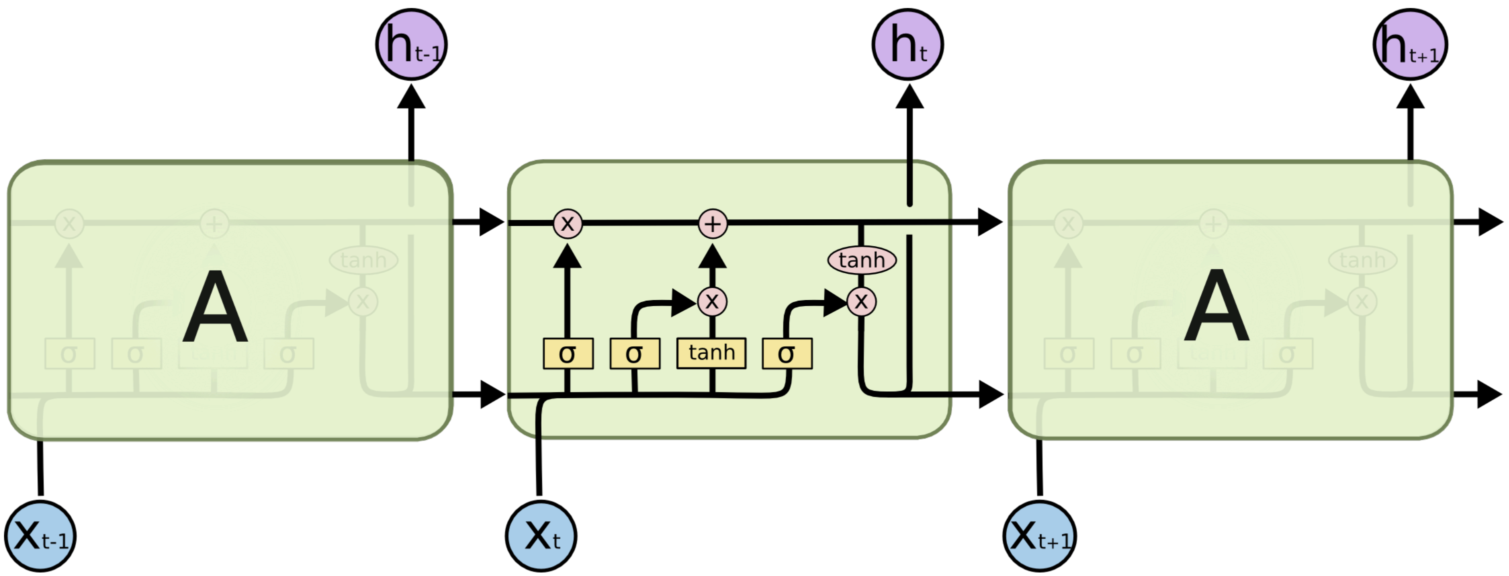
A Representation Learning Framework




R1
......
Rk
Relation Classifier
Contextual Entity Representation
concatenation
( )
Representation Learner
Word Input Text
w(1)
...
w(n-1)
w(n)

Long-Short Term Memory Networks (LSTMs)

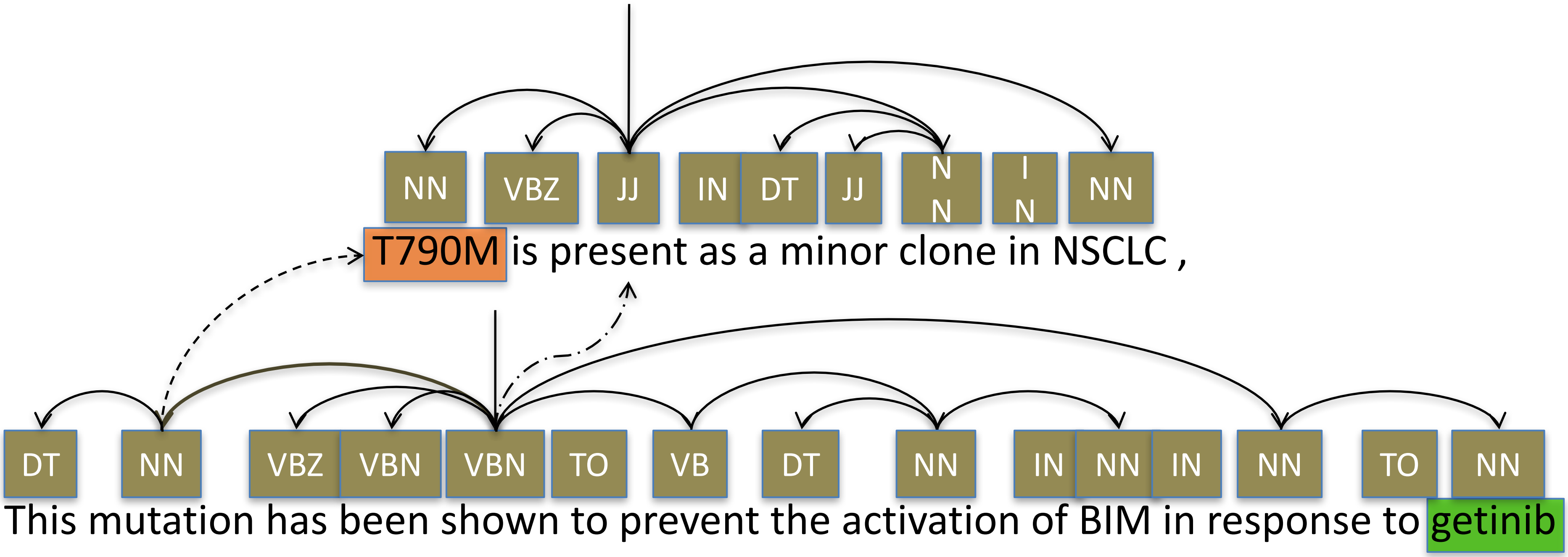
Capture *long-term dependencies* of the input.



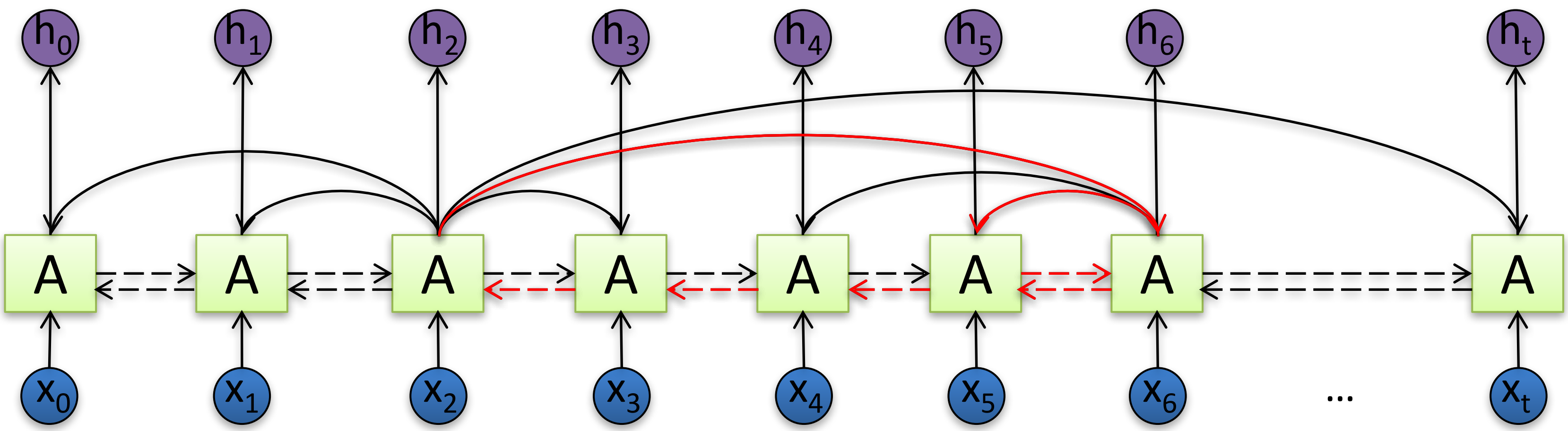
However, it still only explicitly models the dependencies between the adjacent inputs.

Picture credit: colah's blog, 2015

Linguistics Structures for Input Texts



Directed Cyclic Graph

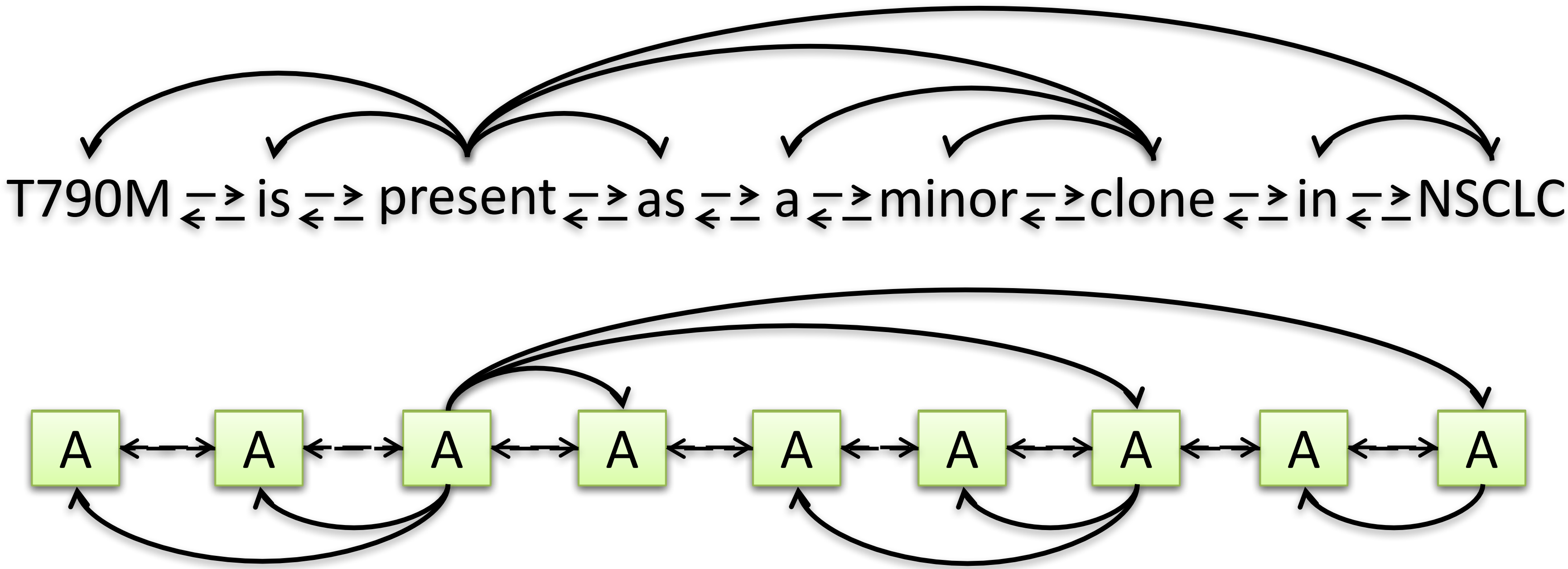


Graph Long Short-Term Memory Networks (Graph LSTMs)

- Goals:
  - *different types* of dependencies: adjacency, *syntactic* dependencies, *coreferences*, and *discourse* relations.
  - *long-distance* dependencies: entities span sentences.
- Challenges: how to define a neural architecture over a cyclic graph?

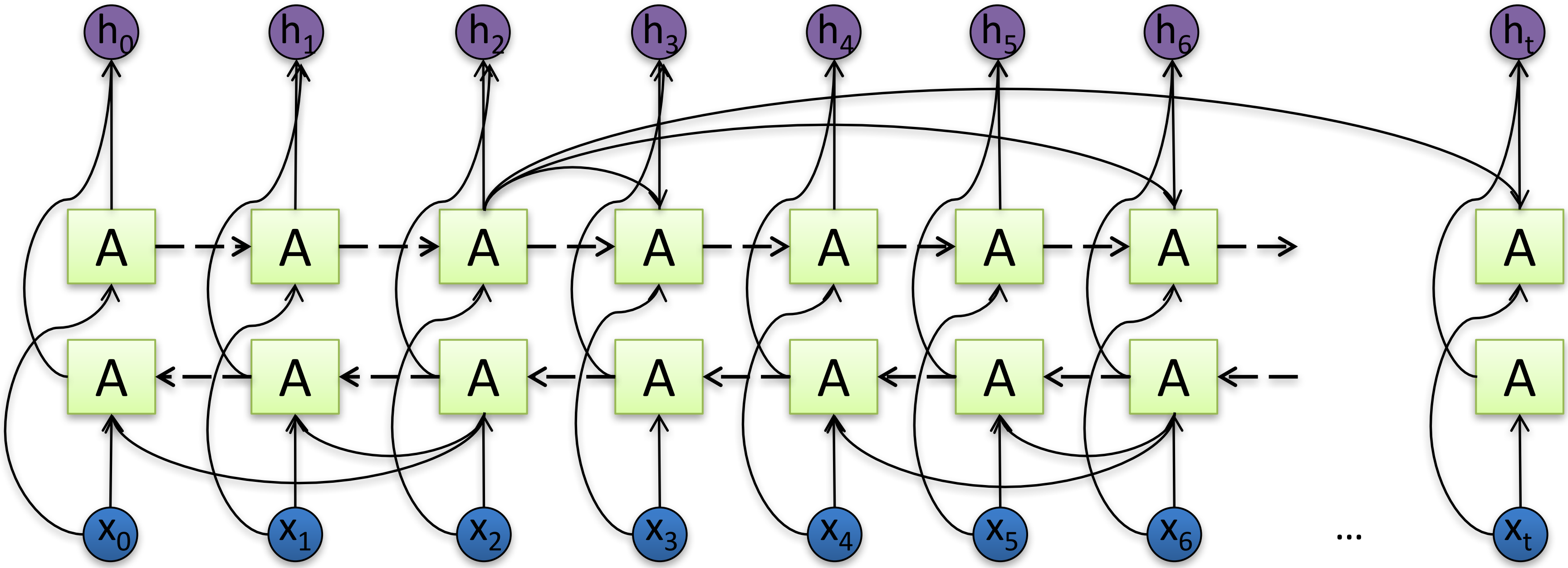
Training Graph LSTMs

- *Provably*, all directed cyclic graph without self-loop can be decomposed into two DAGs.



Training Graph LSTMs

- Approximate a cyclic graph by two directed acyclic graphs (DAGs), and stack the DAGs.



Topological order is well-defined, back propagation training

Chain LSTMs v.s. Graph LSTMs

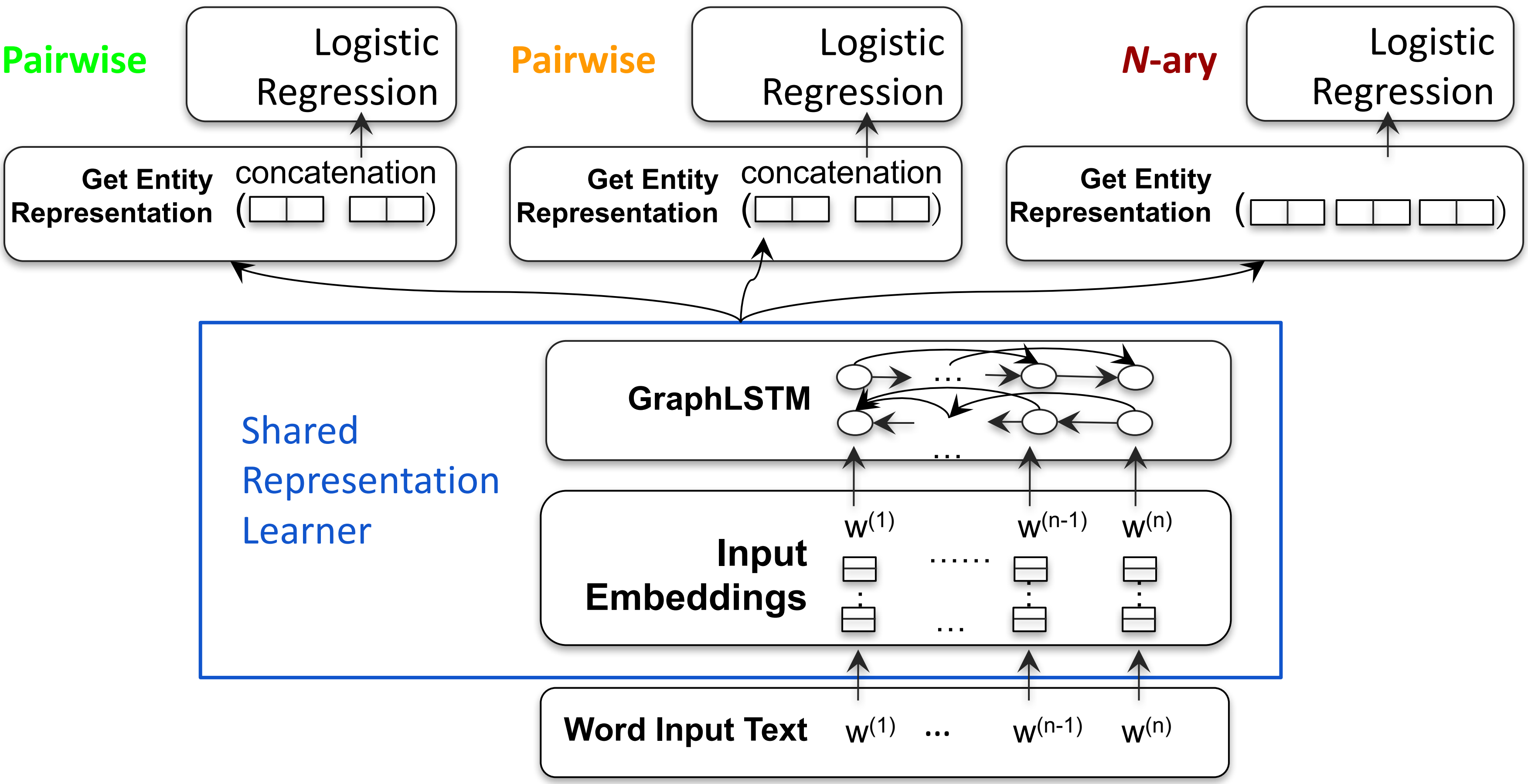
Linear-chain LSTM

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$\boxed{f_t} = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$c_t = i_t \odot \tilde{c}_t + \boxed{f_t \odot c_{t-1}}$$
$$h_t = o_t \odot \tanh(c_t)$$

Graph LSTM (one DAG)

$$i_t = \sigma(W_i x_t + \sum_{j \in P(t)} U_i^{m(t,j)} h_j + b_i)$$
$$o_t = \sigma(W_o x_t + \sum_{j \in P(t)} U_o^{m(t,j)} h_j + b_o)$$
$$\tilde{c}_t = \tanh(W_c x_t + \sum_{j \in P(t)} U_c^{m(t,j)} h_j + b_c)$$
$$\boxed{f_{tj}} = \sigma(W_f x_t + U_f^{m(t,j)} h_j + b_f)$$
$$c_t = i_t \odot \tilde{c}_t + \boxed{\sum_{j \in P(t)} f_{tj} \odot c_j}$$
$$h_t = o_t \odot \tanh(c_t)$$

Multi-task Learning



Domain: Molecular Tumor Board

- Ternary interaction: (drug, gene, mutation)
- Distant supervision
  - Knowledge bases: GDKD + CIVIC
  - Text: PubMed Central articles (~ 1 million full-text articles)
- We got 3,462 paragraphs about drug-gene-mutation relations from distant supervision.

Evaluation of Distant Supervision Relation Extraction is Hard

- 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)
- We can approximate precision
  - Draw a random sample of relations from output, check precision manually
- No way to evaluate recall. Instead, we do absolute recall

Absolute Recall

	Drug	Gene	Mutation	Interaction
DGKD + CiViC	16	12	41	59
Single-Sent	68	228	221	530
Cross-Sent	103	512	445	1461

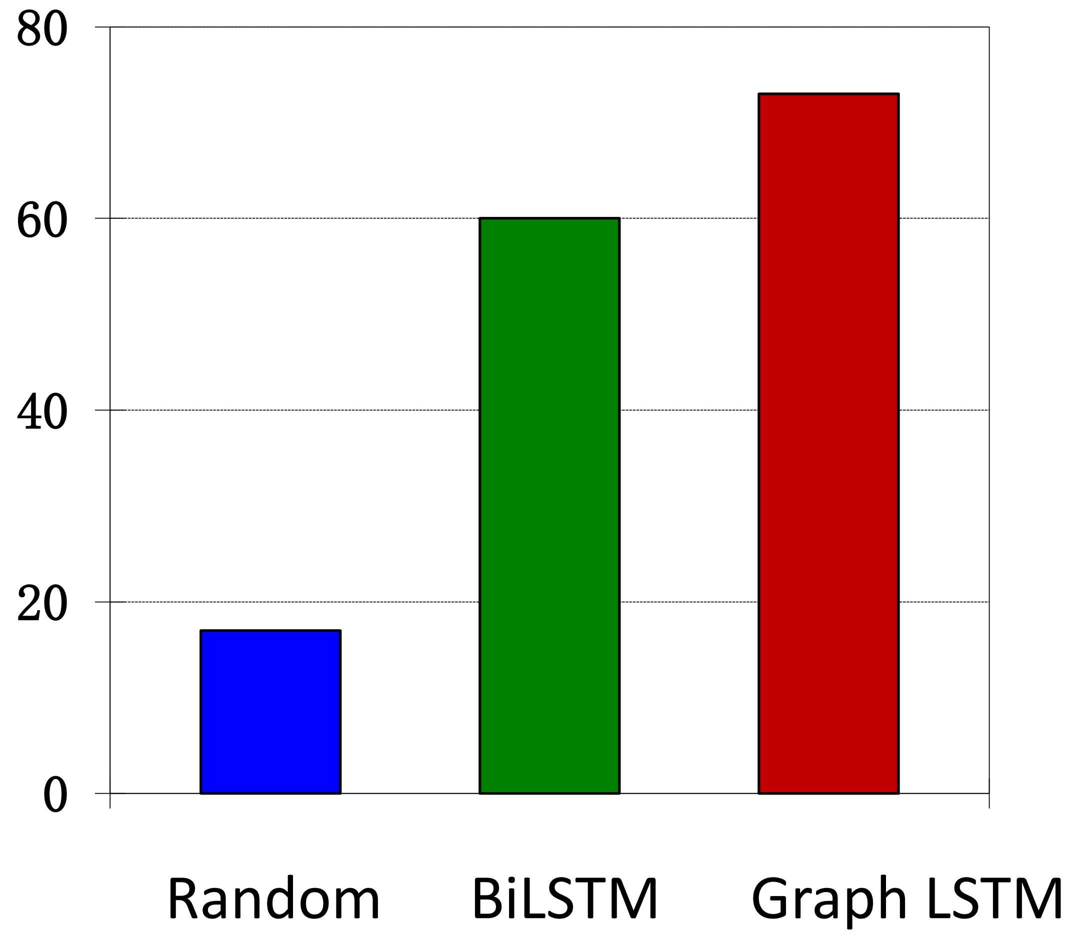
Numbers of *distinct* drugs, genes and mutations and their interactions in the knowledge bases vs. PubMed scale automatic extraction.

Machine reading extracted orders of magnitudes more knowledge

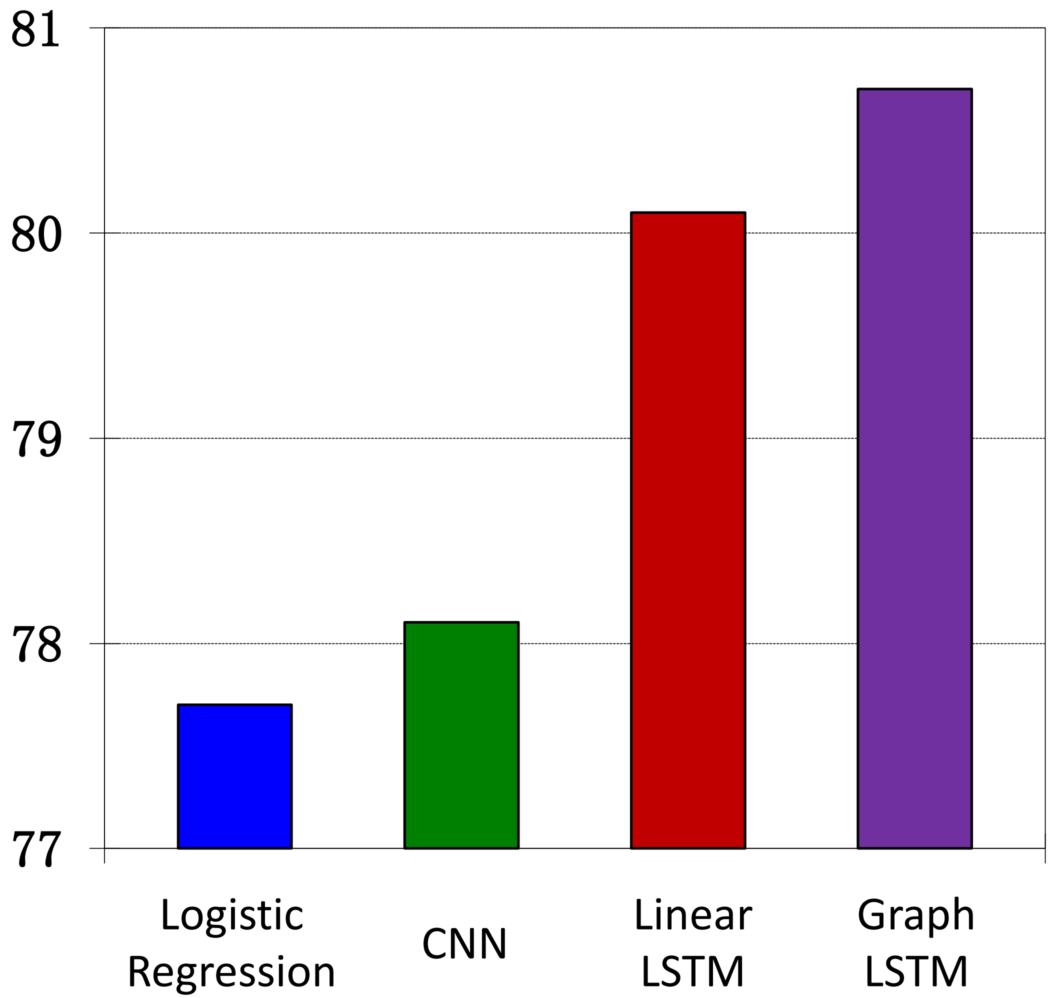
Cross-sentence extraction triples the yield

Sample Precision

Precision



Automatic Evaluation



Multi-Task Learning (Automatic Eval)

Code and data available at: http://hanover.azurewebsites.net/

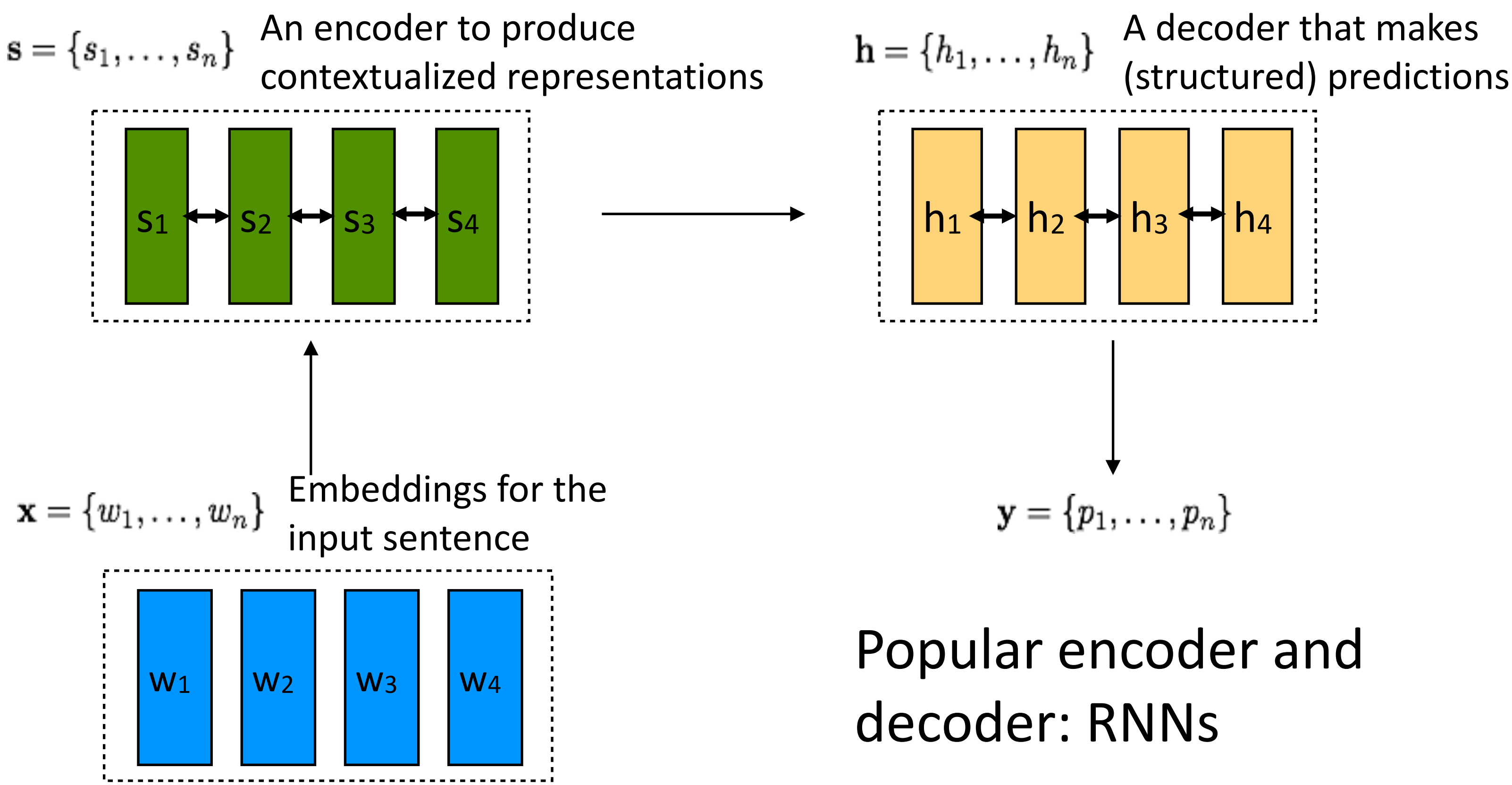
	Drug-Gene-Mutation	Drug-Mutation
Graph LSTM	80.7	76.7
+ Multi-task	82.0	78.5

More results please see Peng et. al. (TACL2017)

Building Robust Models For Low-Resource Settings

- Cross-Sentence N-ary Relation Extraction for Biomedical Domain (low resource domain)
- On Difficulties of Cross-lingual Transfer (low resource languages)
- Plan-and-Write Story Generation (low resource task)

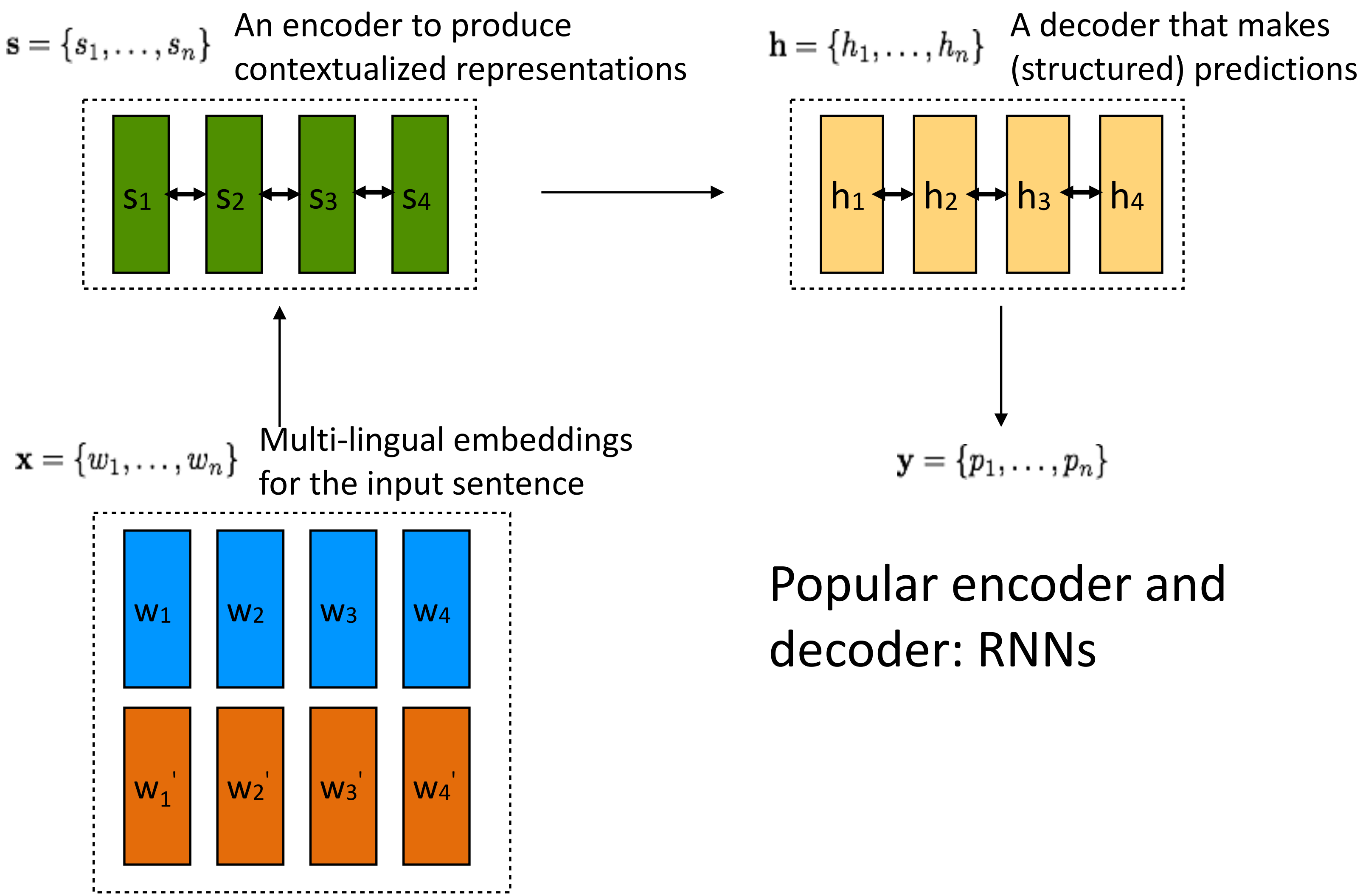
Standard Neural Architectures for NLP



Popular encoder and decoder: RNNs

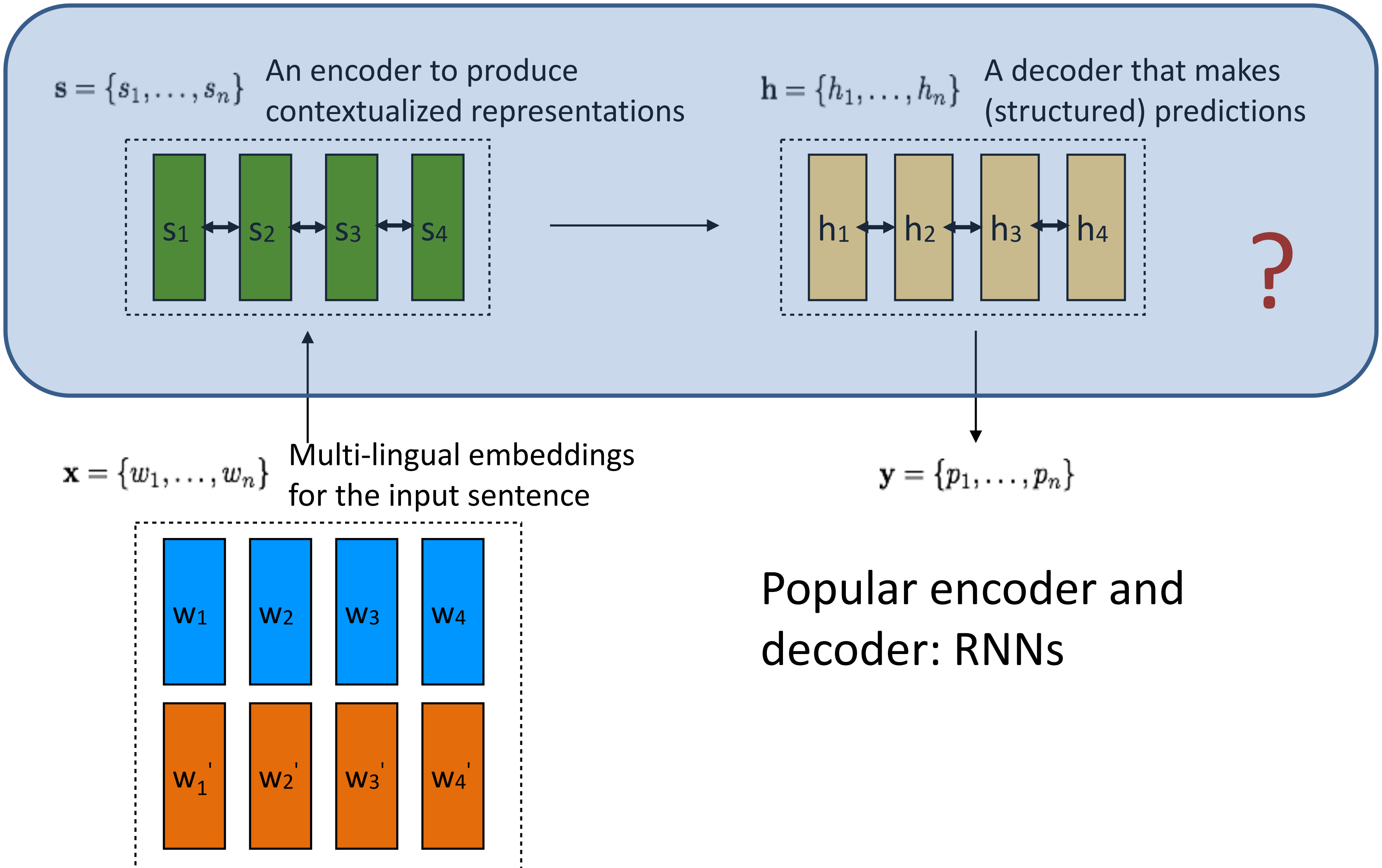
Ahmad et. al. 2018

Cross-Lingual Transfer Learning



Popular encoder and decoder: RNNs

Cross-Lingual Transfer Learning



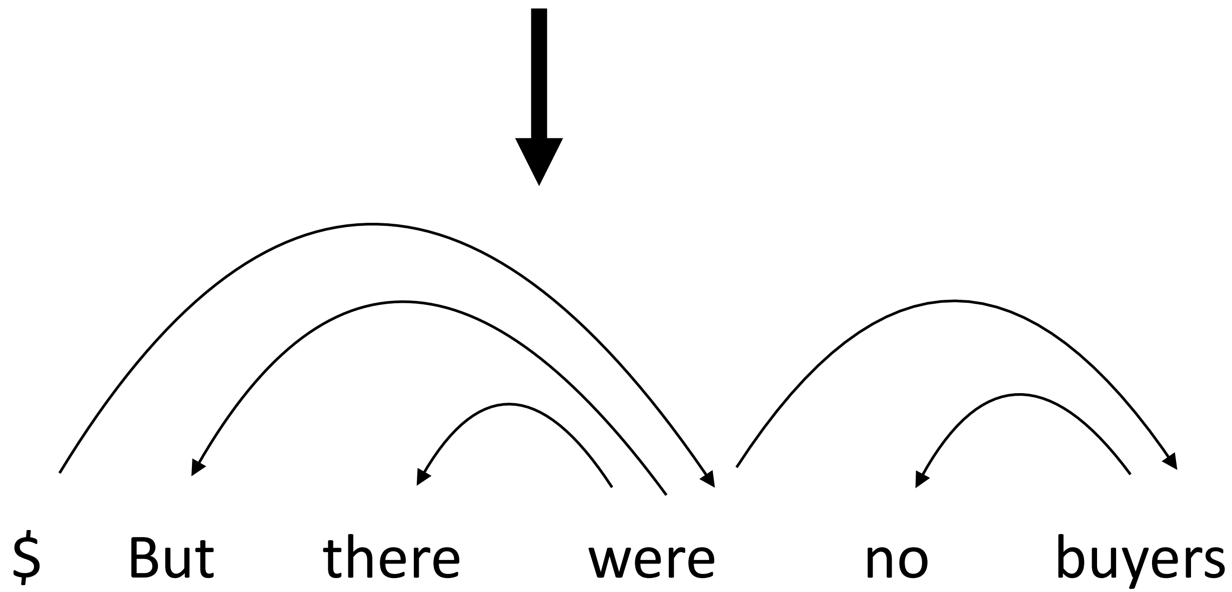
Popular encoder and decoder: RNNs

Are RNNs Good Encoders/Decoders for Cross-lingual Transfer?

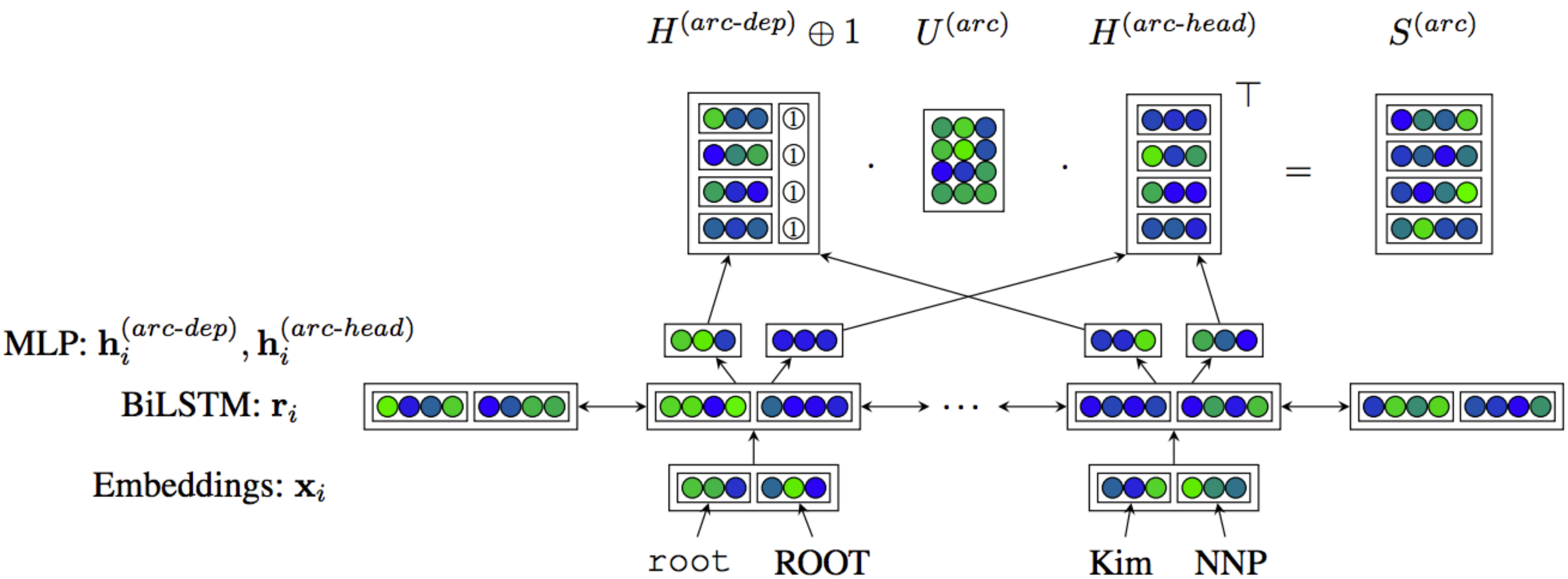
- Overfitting to language-specific order information (our hypothesis)
- Verify an examine our hypothesis on cross-lingual dependency parsing
  - We have UD annotation for over 70 languages
  - Parser is a bottom-level task, directly reflect the problems

Background: Dependency Parsing

$ But there were no buyers



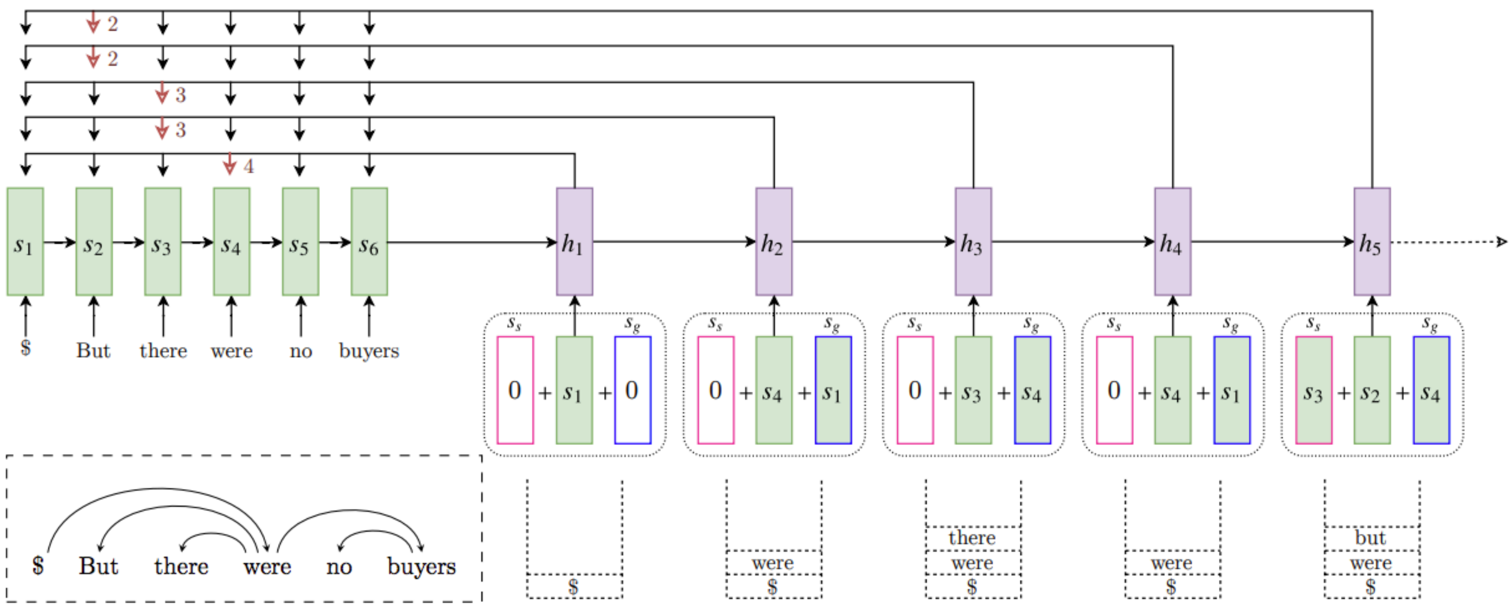
Background: Deep Biaffine Parser



- **Graph-based** parser
- Encoder: Order-sensitive; Decoder: Order-free

Dozat and Manning (ICLR2017)

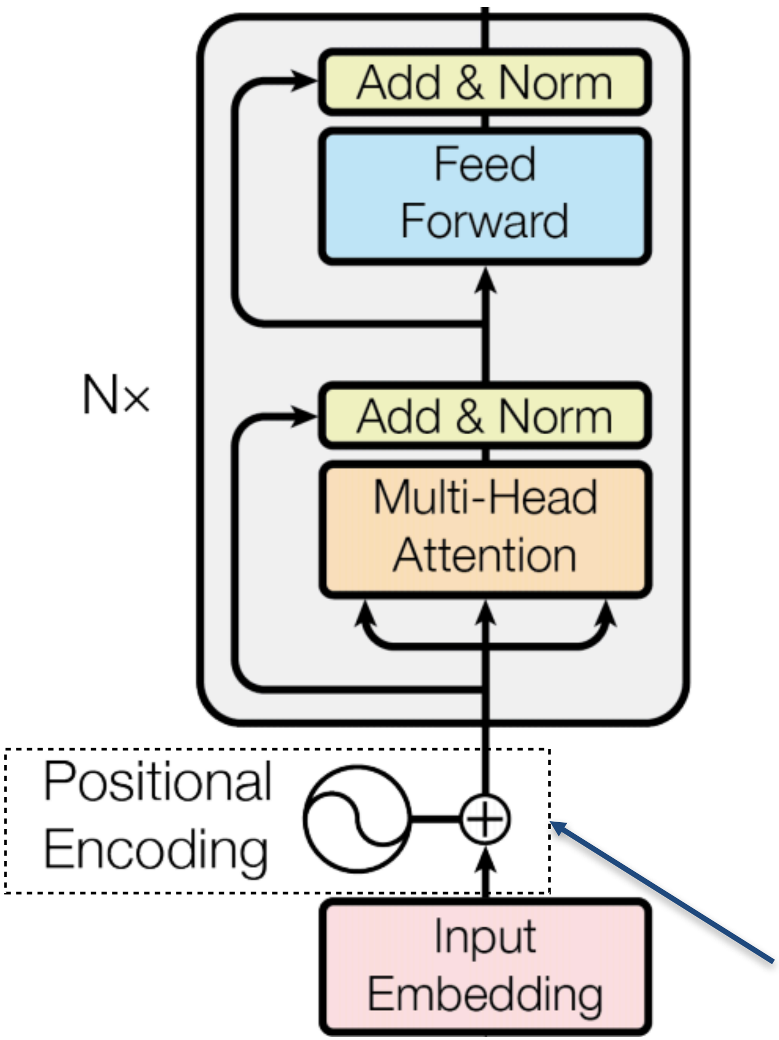
Background: Stack-pointer Networks (StackPtr) Dependency Parsing



- **Transition-based**
- **Order:** Top-down, depth-first
- Actions: "Point" to the next word to choose as a child
- Encoder: Order-sensitive; Decoder: Order-dependent

Ma et. al. (ACL2018)

Background: Multi-Head Self-Attention



Flexible positional encoding (order-free)

- In the original paper:

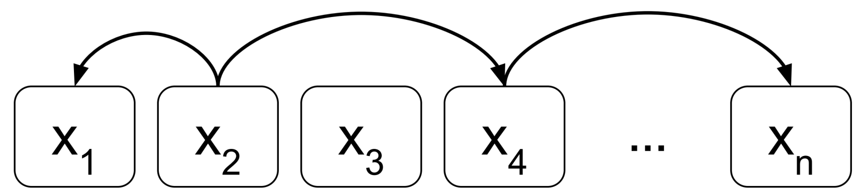
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

Vaswani et. al. (NIPS 2017)

- Relative positional embeddings:

$a^V_{2,1}=w^V_{-1}$ $a^V_{2,4}=w^V_{2}$ $a^V_{4,n}=w^V_{k}$

$a^K_{2,1}=w^K_{-1}$ $a^K_{2,4}=w^K_{2}$ $a^K_{4,n}=w^K_{k}$

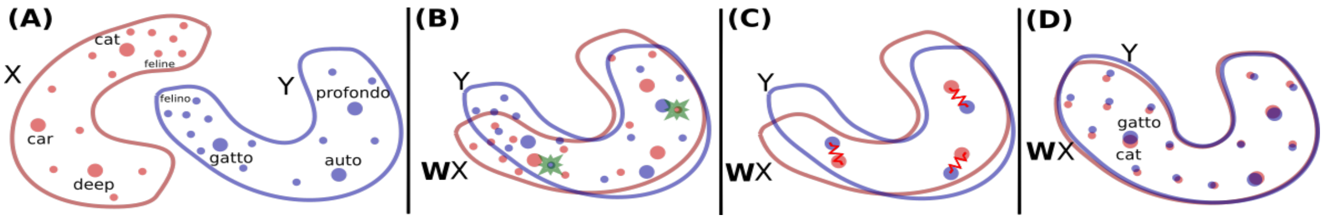


Shaw et. al. (NAACL2018)

Architectures for Cross-lingual Parser

- Embedding

Facebook MUSE



Conneau et. al. ICLR2018

- Encoders
  - BiLSTMs (order-sensitive) v.s.
  - Multi-Head Self-Attention (order-free)
- Decoders
  - Pointer Network (order-sensitive) v.s.
  - BiAffine Attention (order-free)

Experiments

- Datasets:
  - Universal Dependency Treebanks (V2.2)
  - 31 languages, 12 families
- Setting:
  - Train and develop on English
  - Directly predict on the rest 30 languages (zero-shot)

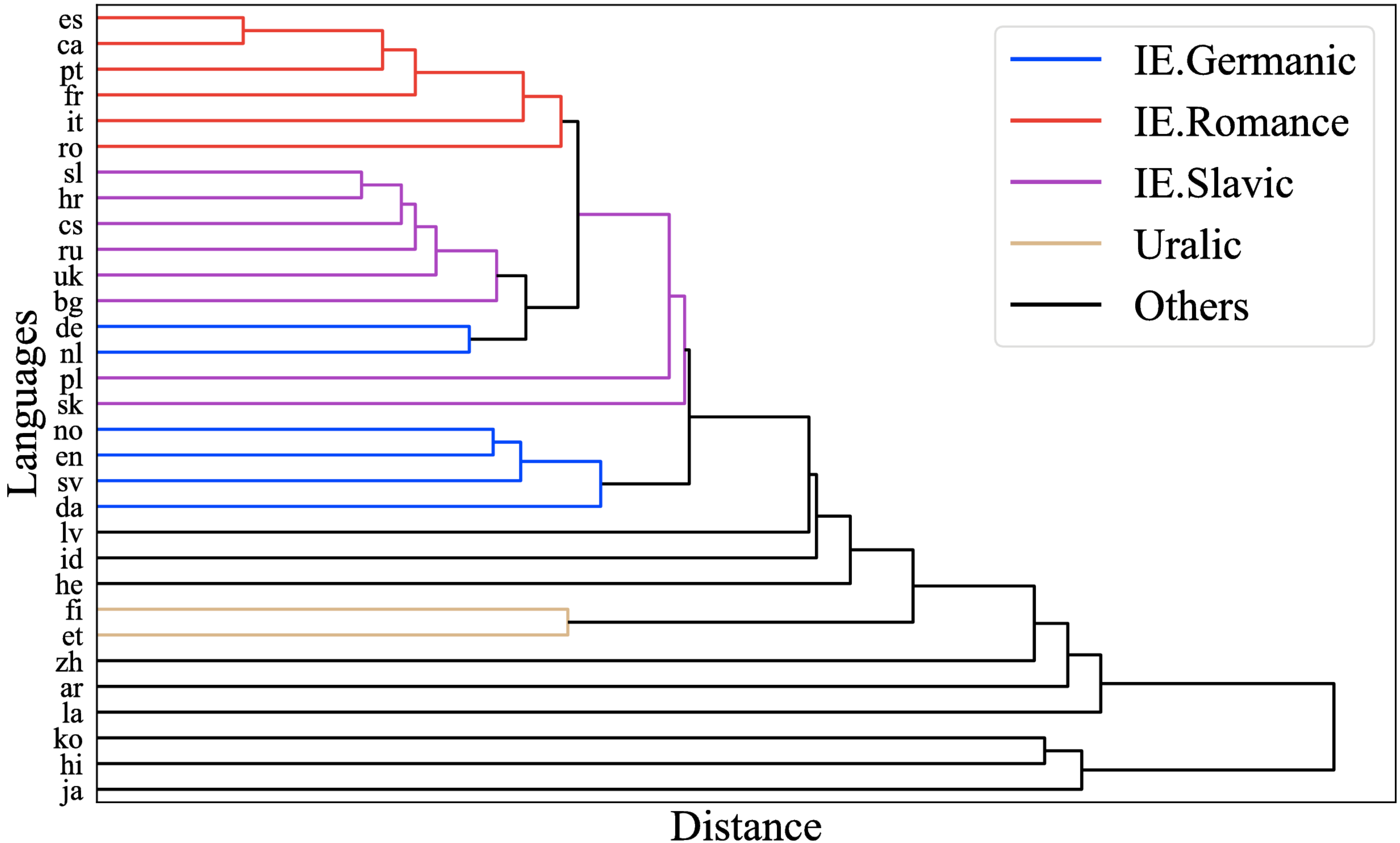
Datasets Details

Language Families	Languages
Afro-Asiatic	Arabic (ar), Hebrew (he)
Austronesian	Indonesian (id)
IE.Baltic	Latvian (lv)
IE.Germanic	Danish (da), Dutch (nl), English (en), German (de), Norwegian (no), Swedish (sv)
IE.Indic	Hindi (hi)
IE.Latin	Latin (la)
IE.Romance	Catalan (ca), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Spanish (es)
IE.Slavic	Bulgarian (bg), Croatian (hr), Czech (cs), Polish (pl), Russian (ru), Slovak (sk), Slovenian (sl), Ukrainian (uk)
Japanese	Japanese (ja)
Korean	Korean (ko)
Sino-Tibetan	Chinese (zh)
Uralic	Estonian (et), Finnish (fi)

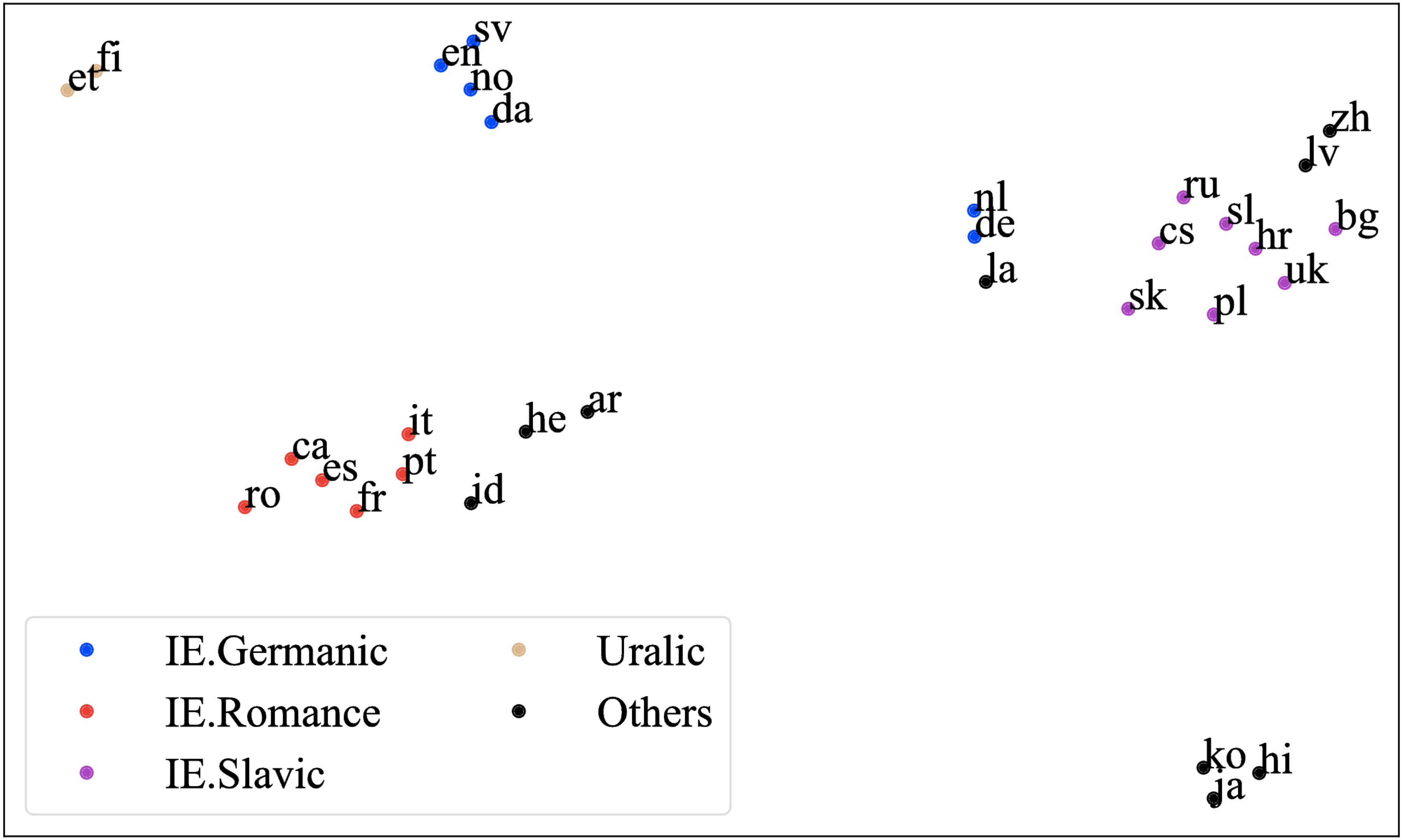
Characterizing Language Distances

- Augmented dependency label features:
  - Triple (ModifierPOS, HeadPOS, DependencyLabel), e.g. (PRON, VERB, obj)
  - Feature selection: exists in > 24 languages and with > 0.1% relative frequency
  - Feature value: left (modifier before head) frequency and right (modifier after head) frequency
  - 52 feature types (104 features) total.

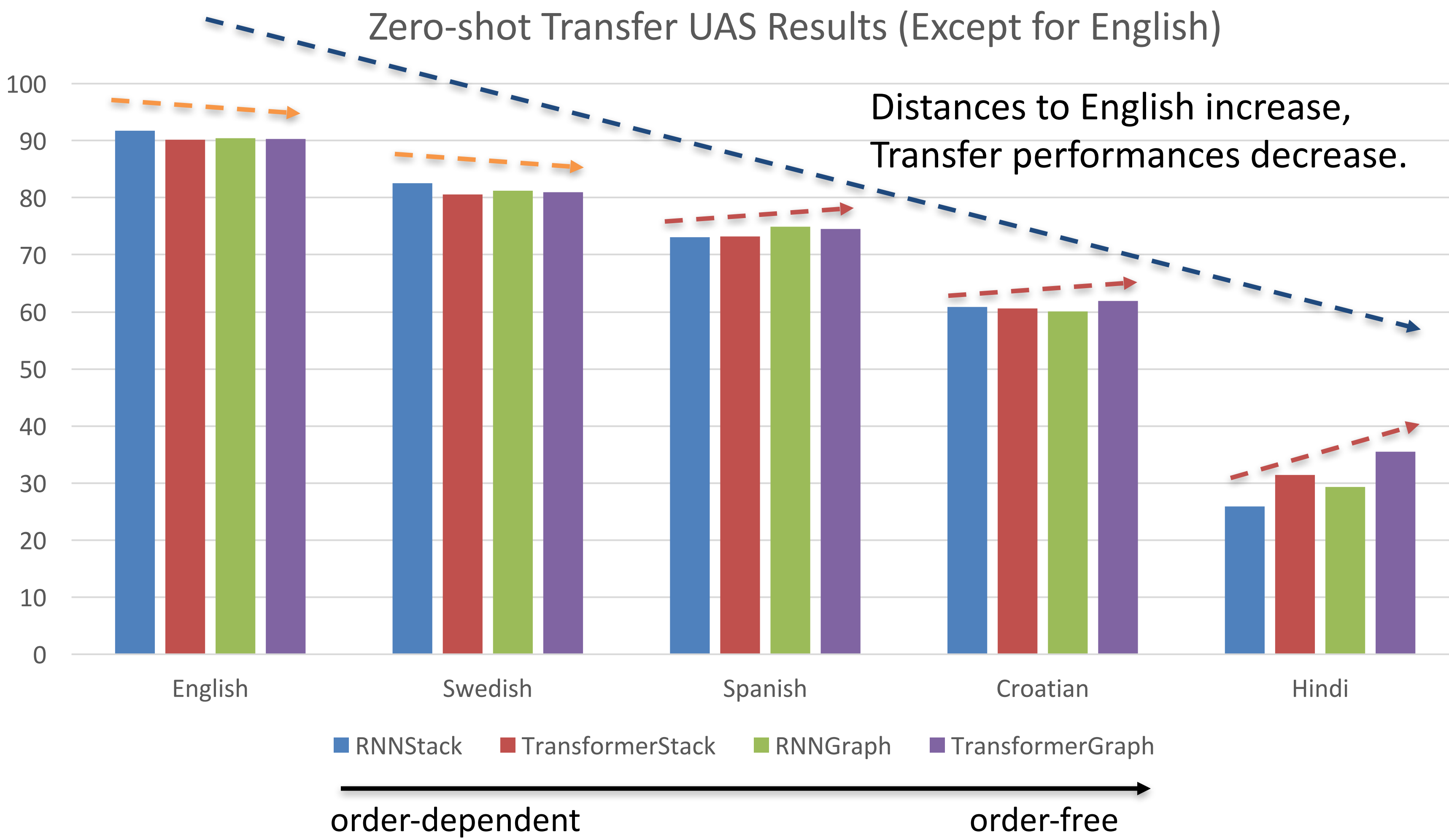
Word Order Characterizes Language Distances



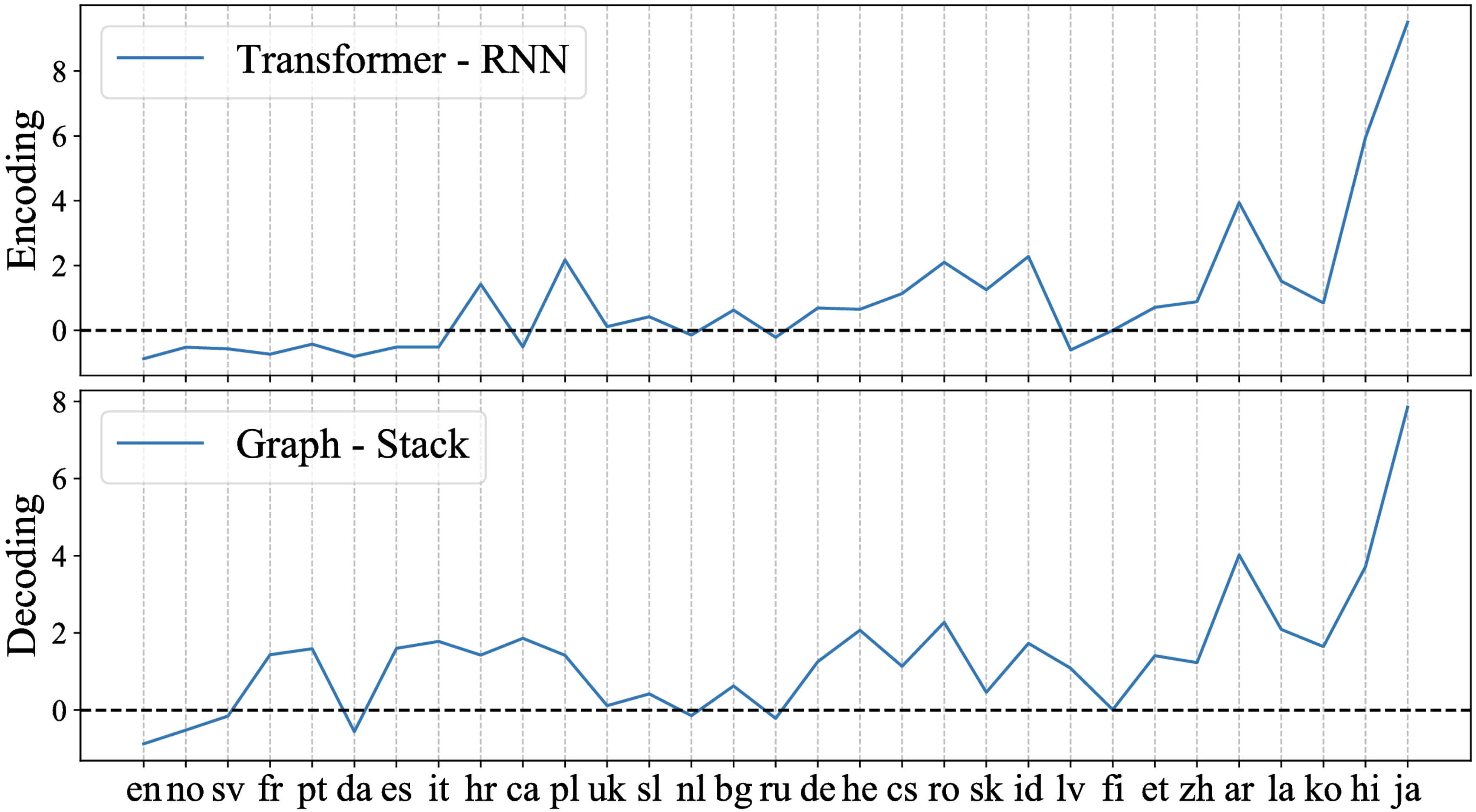
Word Order Characterizes Language Distances



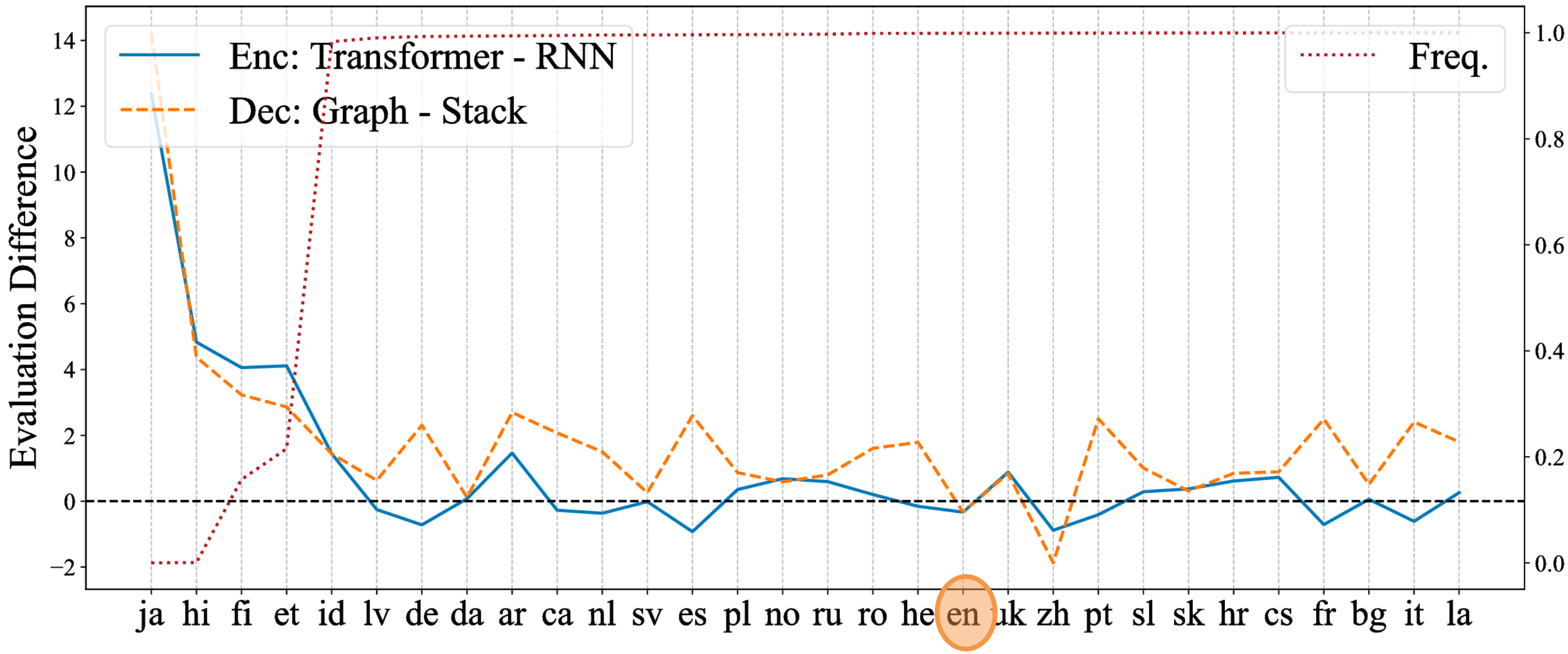
Selected Transfer Results of Different Architectures



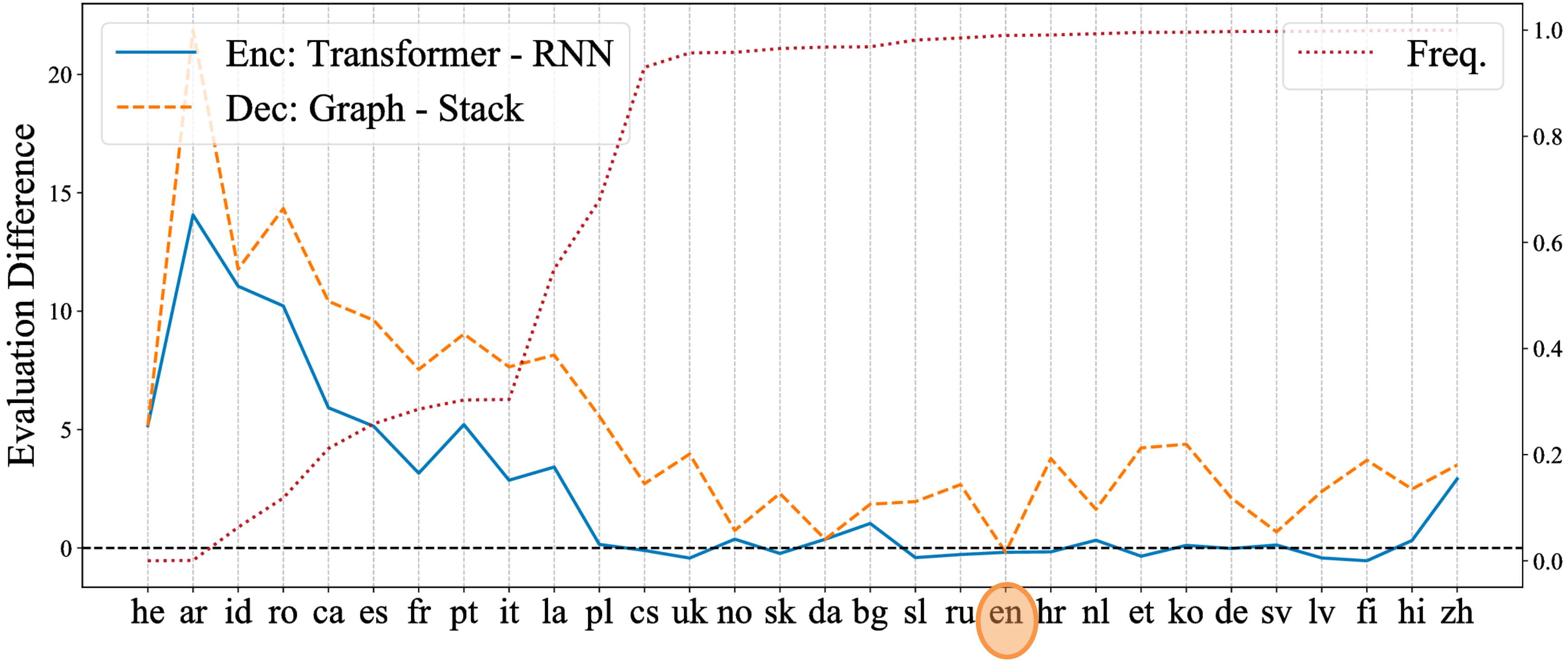
Overall Comparisons of Order-Free v.s. Order-Dependent Encoders/Decoders



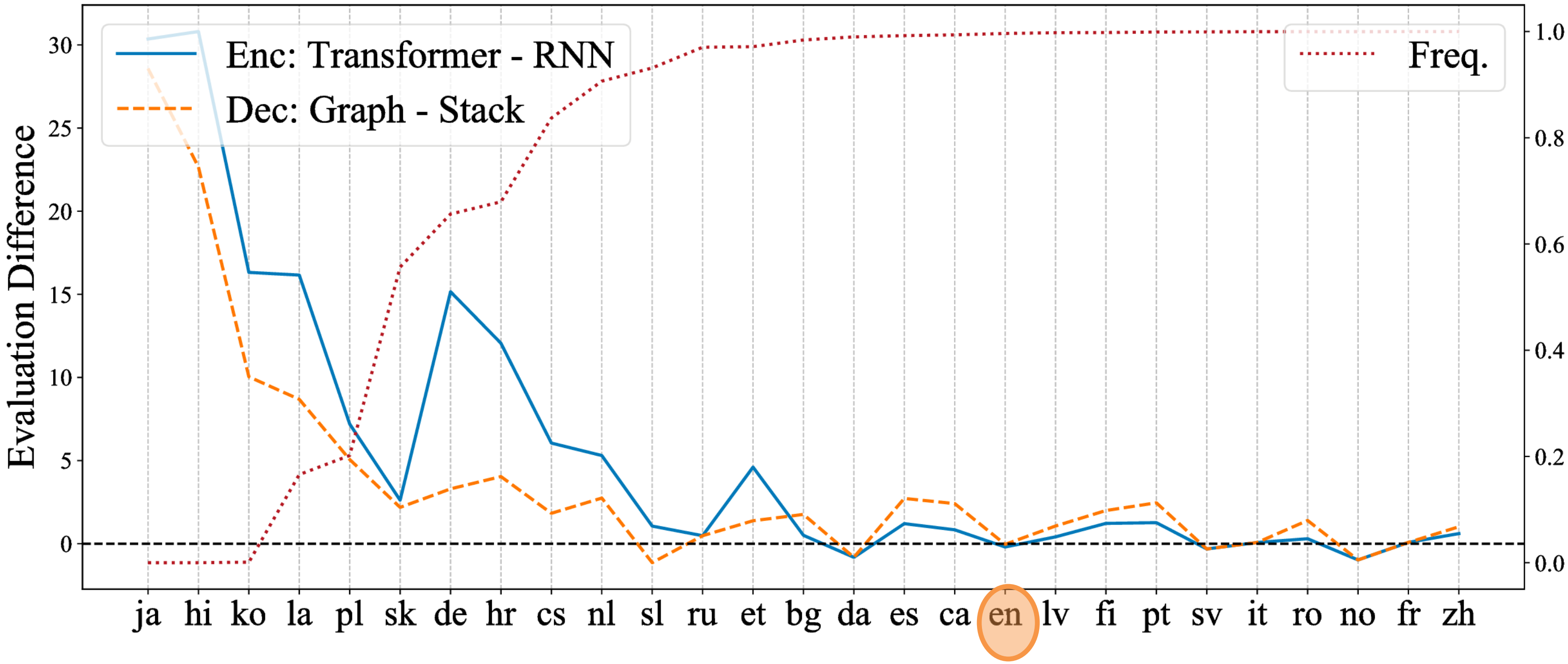
Case Study -- (ADP, NOUN, case)



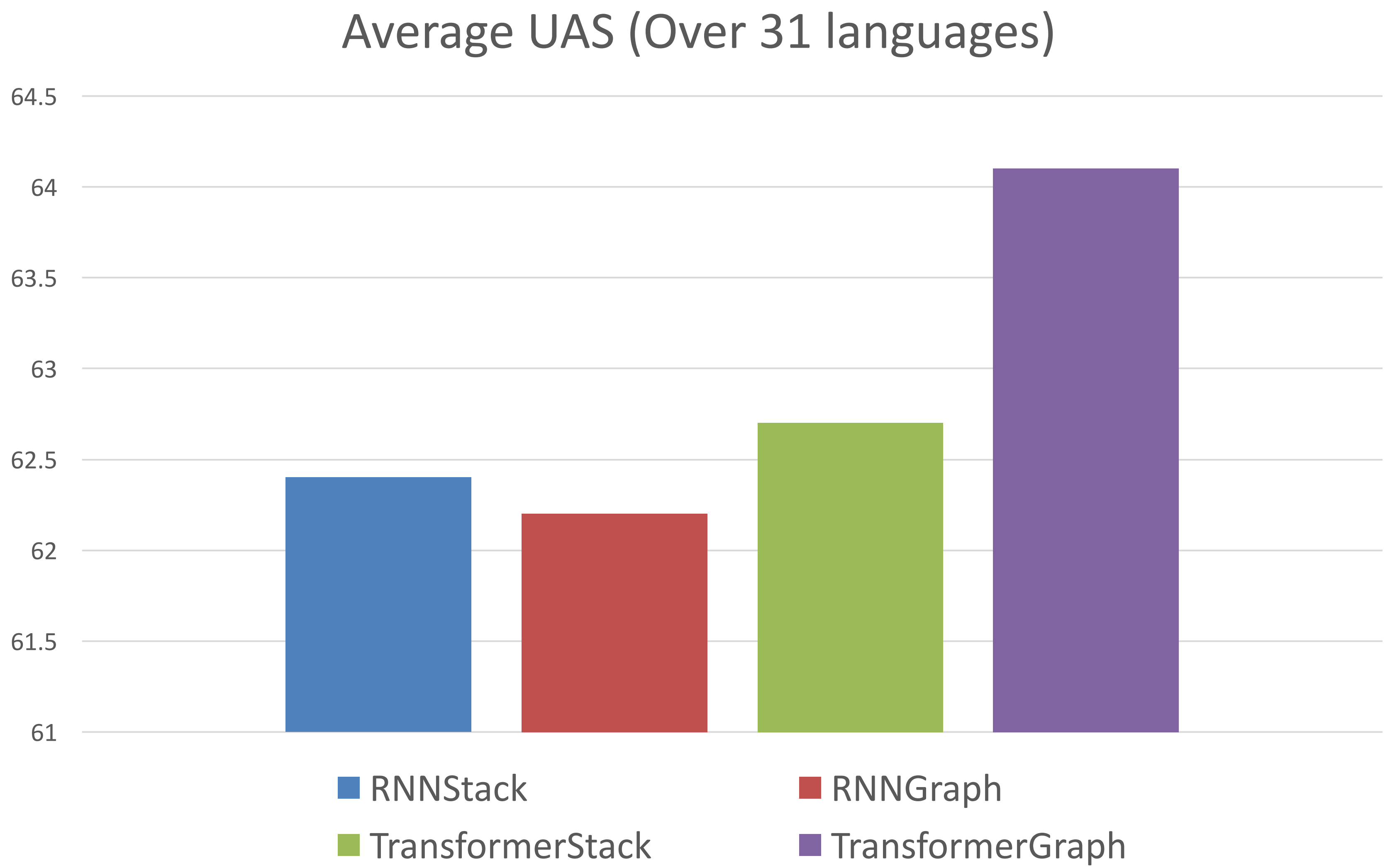
Case Study -- (ADJ, NOUN, amod)



Case Study -- (AUX, VERB, aux)



Overall Performances

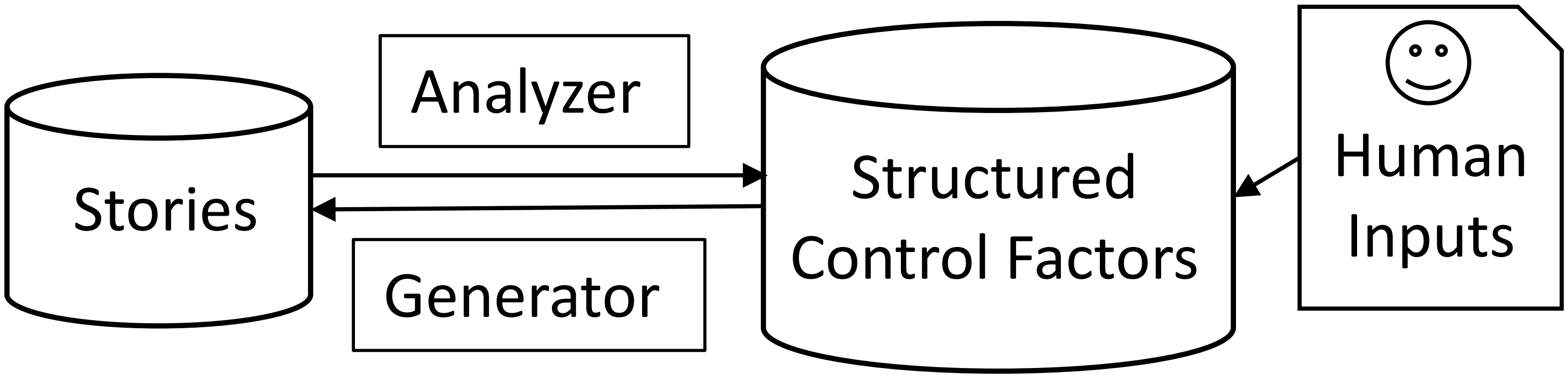


Building Robust Models For Low-Resource Settings

- Cross-Sentence N-ary Relation Extraction for Biomedical Domain (low resource domain)
- On Difficulties of Cross-lingual Transfer (low resource languages)
- Plan-and-Write Story Generation (low resource task)

Story Generation

- What are in a story?
  - Characters, key events, morals, conflicts, sentiment…
- We want to incorporate all the aspects
  - Unfortunately, even human do not have clear understanding about what's in a story. There are few annotations.
- Analyzing stories to generate stories with minimal or no supervision.



Yao & Peng et. al. (AAAI2019)

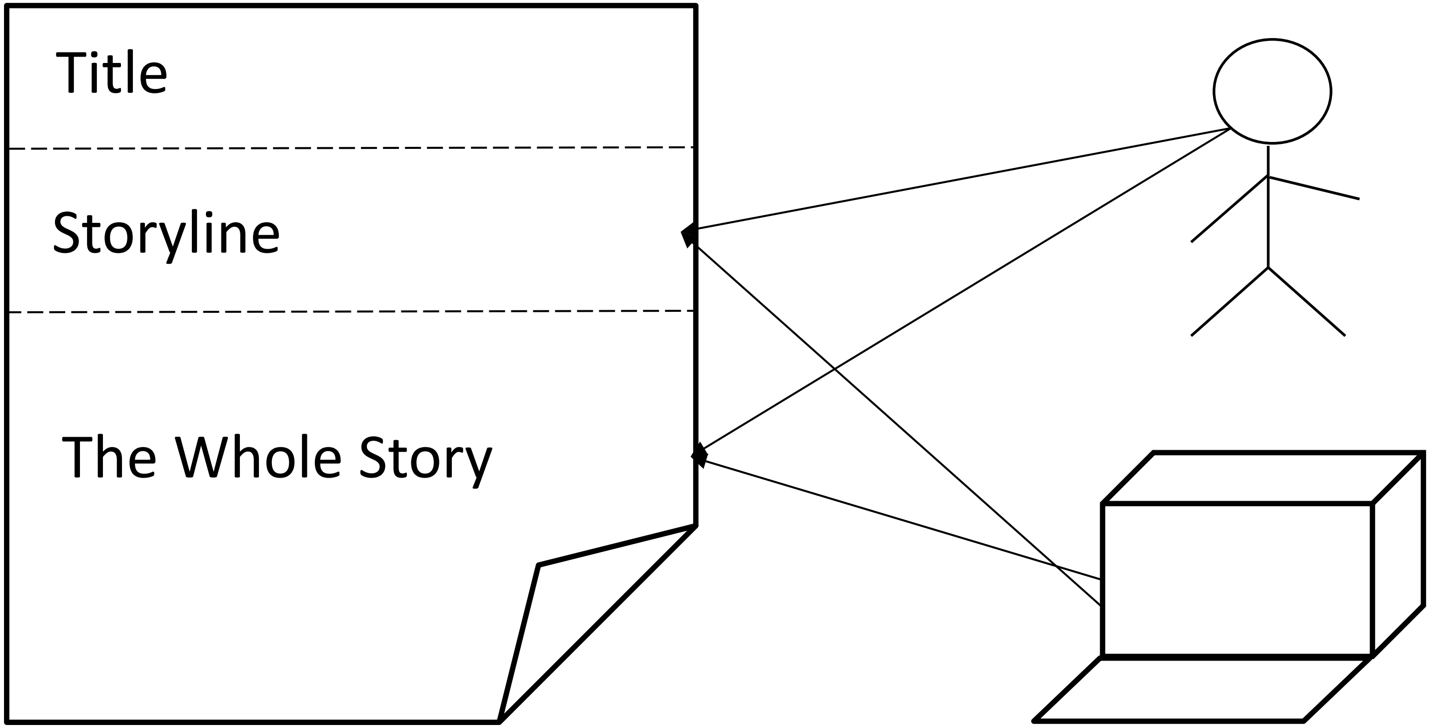
Problem of (Neural) Story Generation

- **Title:** bicycle path accident
- **Generated Story:** sam bought a new bicycle. his bicycle was in an accident. his bicycle was in an accident. his bicycle was in an accident. his bicycle was totaled.
- **Title:** darth vader on earth
- **Generated Story:** it was a very windy day. i 've never been to it before. i do not know what to do. i do not know what to do. i think it is a good idea.

Plan-and-Write Hierarchical Generation

- Can computer generate storylines automatically (given titles)?
  - Equip our system with the ability to model "what happens next".
  - Mimic human writers' common practice of writing sketches: have a big picture.
  - Computer and human can interactively modify the storylines, more fun interactions.

Interactive Generation Task



Extracting Storylines

- The ROCStories dataset: 98,161 turker-written five-line stories with titles.
- Extract one word or phrase from one sentence using RAKE algorithm proposed in the IR community (2010).
- Use the word/phrase sequence as an approximation of the storyline.

Examples

Title: christmas shopping

Story: frankie had christmas shopping to do.

she went to the store.

inside, she walked around looking for gifts.

soon her cart was full.

she paid and took her things home.

Storyline (unsupervised extraction): frankie store gifts cart paid

Title: farm

Story: bogart lived on a farm.

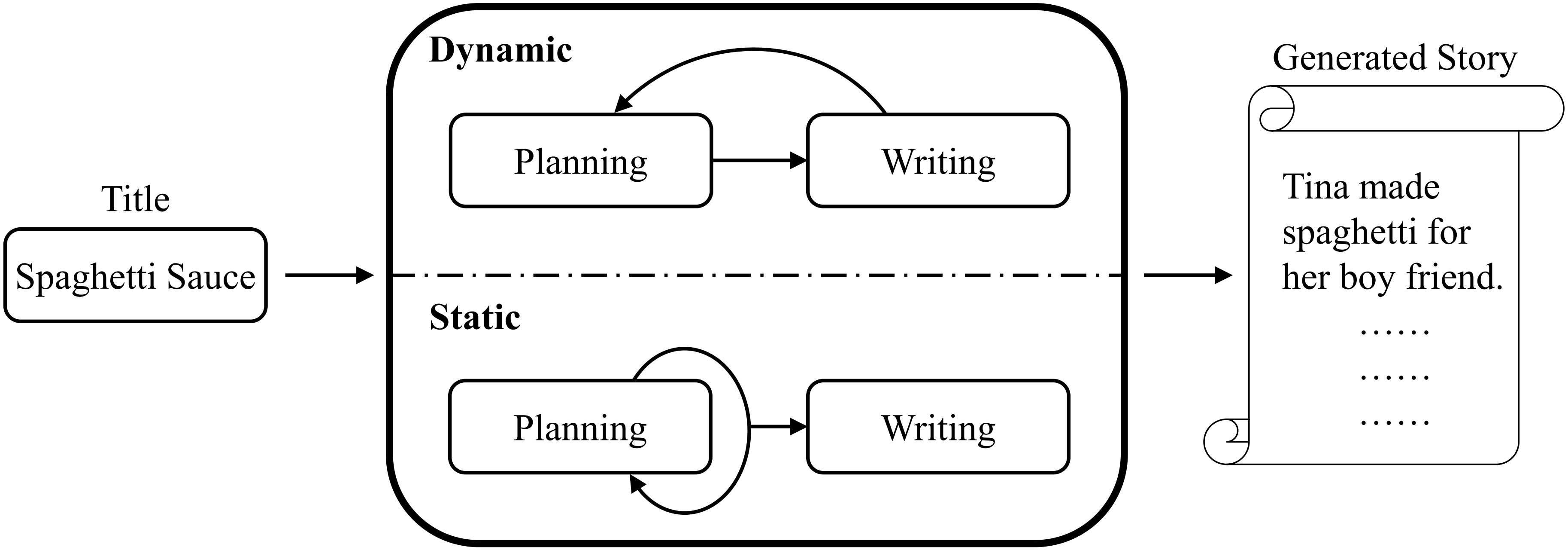
he loved bacon.

he decided to buy a pig. shortly after, he grew fond of the pig.

bogart stopped eating bacon.

Storyline (unsupervised extraction): farm bacon decided pig bogart

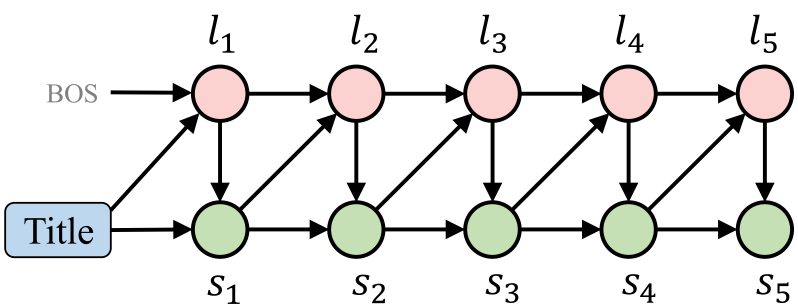
Plan-and-Write Overview



The *planning* component generates storylines from titles. The *writing* component generates stories from storylines and titles.

Dynamic and Static Schemas

Dynamic Schema



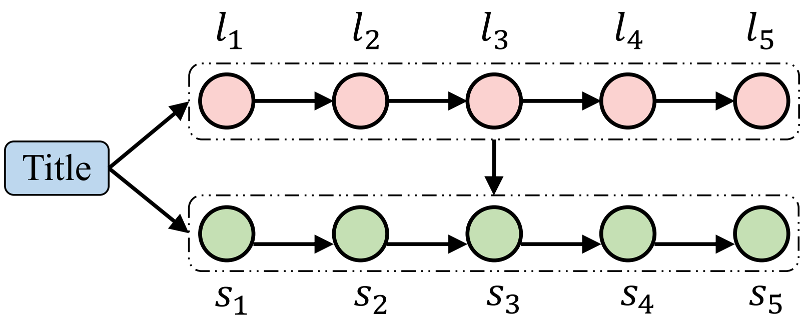
We define context as: $\boldsymbol{ctx} = [\boldsymbol{t}; \boldsymbol{s}_{1:i-1}]$

At the plan step, we model: $P(li|\textbf{ctx}, l_{1:i-1})$

At the write step, we model: $P(s_i|\textbf{ctx}, l_{1:i})$

The probabilities are computed by some specifically designed fusion-RNN cells.

Static Schema



At the plan step, we model: $P(li|\textbf{t}, li_{-1)}$

At the write step, we model: $P(s_i|\textbf{ctx},l_{1:5})$

The probabilities are computed by standard language models and sequence to sequence with attention models.

Some Observations

- Plan-and-Write strategies generate more interesting, less repetitive stories.
- Plan-and-Write strategies generate more on-topic stories.
- Static strategy works better than dynamic strategy.

Generation Results

Without Storyline Planning

Title: gymnastics

Story (generated):

i wanted to learn how to draw.

so, i decided to go to the gym.

i went to the local gym.

i got a lot of good grades.

i was very happy.

With Storyline Planning

Title: gymnastics

Storyline (generated): wanted decided class practiced well

Story (generated):

i <u>wanted</u> to be a gymnast.

i <u>decided</u> to learn how to do gymnastics.

i decided to take a <u>class</u>.

i <u>practiced</u> every day.

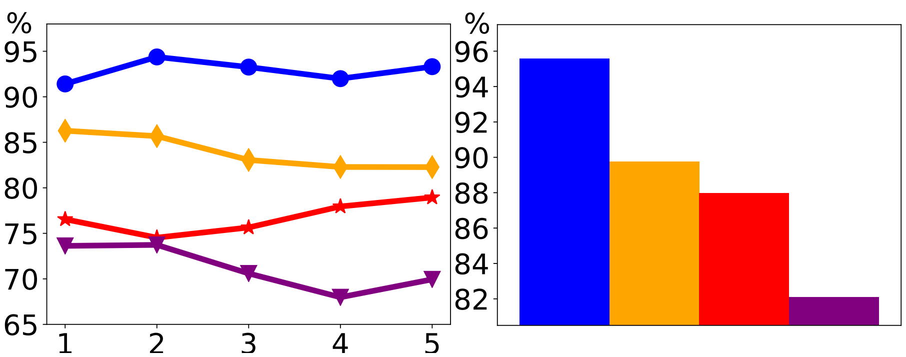
i was able to do <u>well</u> on the class.

Generation Results (Cont.)

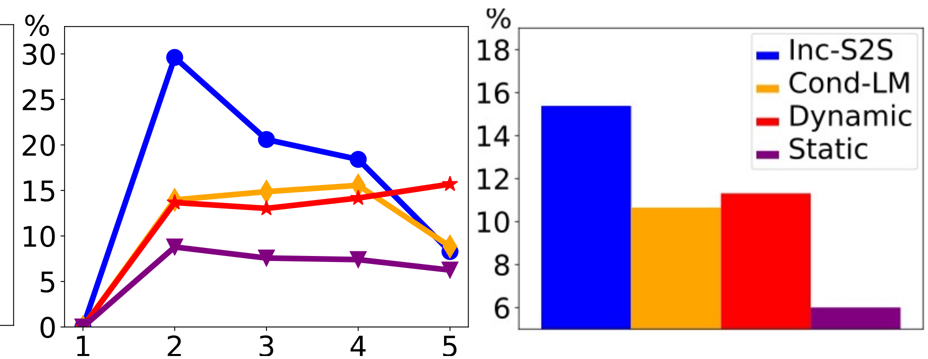
Without Storyline Planning	With Storyline Planning
Title: rock jumping **Story (generated):** i was at the park with my friends. i was playing with my friends. i was playing with my friends. i tripped over a rock. i fell on the ground.	**Title:** rock jumping **Storyline (generated):** day decided jumped fell broke **Story (generated):** one <u>day</u> , i <u>decided</u> to go rock jumping. i <u>jumped</u> and fell. i <u>fell</u> and <u>broke</u> my ankle. i had to go to the hospital. i learned to be more careful next time .

Quantitative Results on Repetition

Inter- and intra-story tri-grams repetition rates by sentences (curves) and for the whole stories (bars), the lower the better. We also conduct the same computation for four and five-grams and observed the same trends. As reference points, the whole story repetition rates on the human-written training data are 34% and 0.3% for the inter- and intra-story measurements respectively.



Inter-story repetition rates



Intra-story repetition rates

User Preferences

Aspect	Dynamic v.s. Inc-S2S			Static v.s. Cond-LM			Static v.s. Dynamic		
	Dyna.	Inc.	Kap.	Static	Cond.	Kap.	Static	Dyna.	Kap.
Fidelity	**35.8%**	12.9%	0.42	**38.5%**	16.3%	0.42	**38.0%**	21.5%	0.30
Coherence	**37.2%**	28.6%	0.30	**39.4%**	32.3%	0.35	**49.5%**	28.3%	0.36
Interesting	**43.5%**	26.7%	0.31	**39.5%**	35.7%	0.42	**43.6%**	34.4%	0.35
Overall	**42.9%**	27.0%	0.34	**40.9%**	34.2%	0.38	**50.1%**	30.1%	0.38

The human study is conducted on Amazon Mechanical Turk (AMT). 233 users were participated in the study.

Analysis

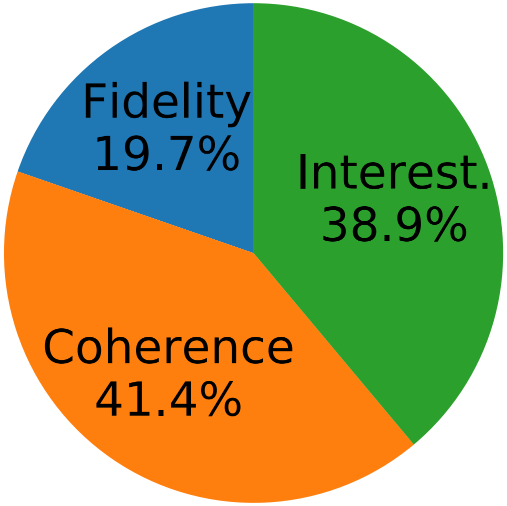


Figure: The regression coefficient that shows which aspect is more important in human evaluation of stories.

Method	B1	B2	L-S
Dynamic	6.46	0.79	0.88
Static	9.53	1.59	0.89

Table: The storyline BLEU score (only BLEU-1 and BLEU-2) and the correlation between storylines and stories.

Yao & Peng et. al. (AAAI2019)

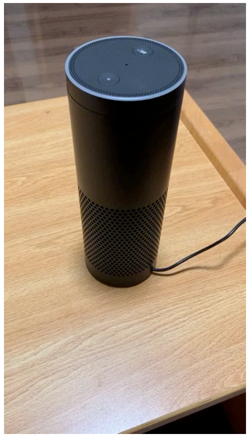
Lingering Problems

Yao & Peng et. al. (AAAI2019)

Title (problem)	Generated Storyline	Generated Story
pulled over (off-topic and strange language)	driving road car hit time	tom was driving to work. he was driving down the road. a car hit him in the car. tom hit the car. he had to get a new car.
saving money (inconsistent)	money wanted decided job pay	i had a lot of money. i wanted to save money. i decided to get a job. i got a job at a local bank. i was able to pay my bills.
my nephew's vacation (off-topic and inconsistent)	family vacation trip decided time	i went on vacation with my family. we went on a vacation. we had a great time. we decided to go on a trip. we had a great time.

Summary

- Low-resource settings are everywhere in NLP
  - Low-resource languages
  - Low-resource domains
  - Low-resource tasks
- Need more robust/generalizable and intelligent NLP and machine learning algorithms
- Creative composition is a fun and challenging task!
- Alexa storytelling demo



Thanks!
Questions?

[Demo: http://cwc-story.isi.edu/](http://cwc-story.isi.edu/)