Document Modeling and Discourse Analysis

Lili Mou
doublepower.mou@gmail.com

Document Modeling with Gated Recurrent Neural Network for Sentiment Classification

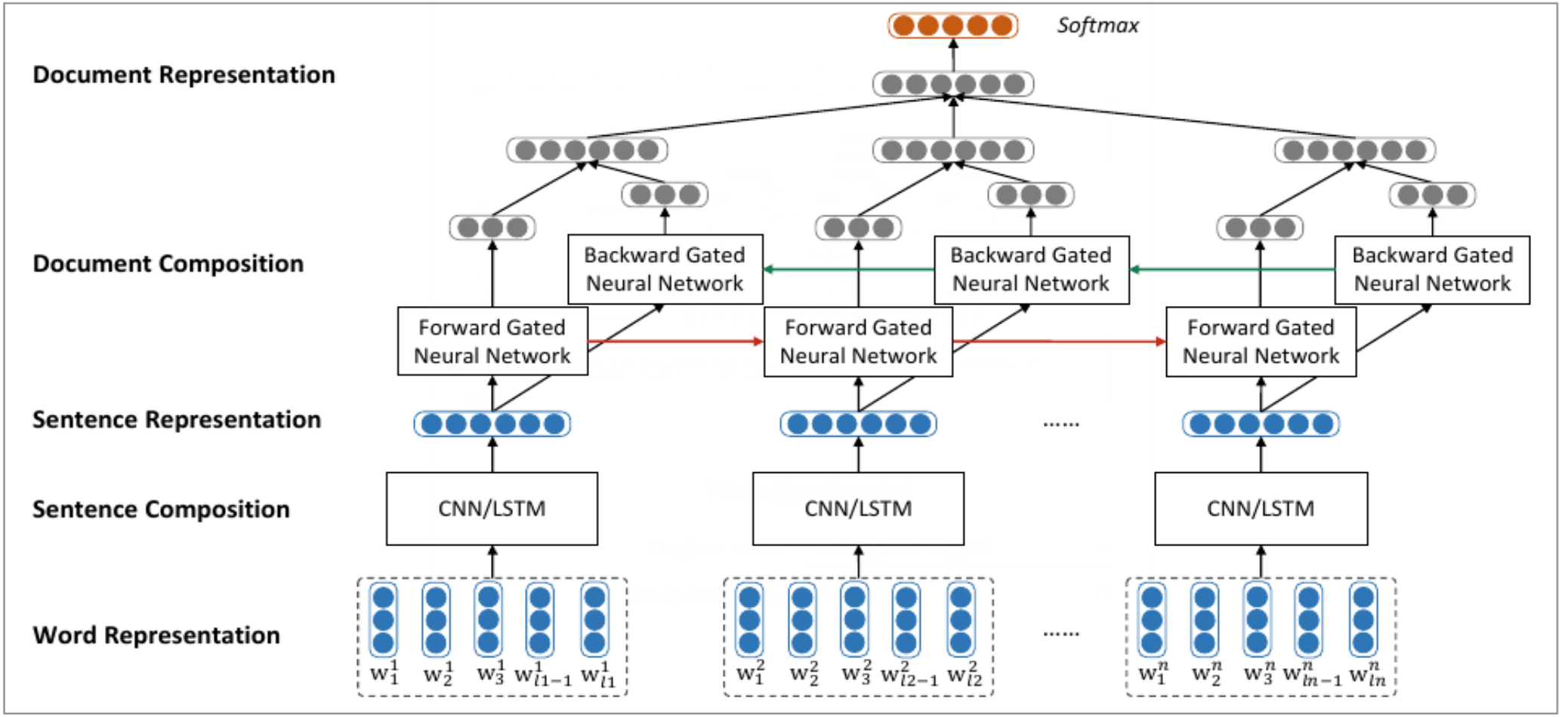
LONG

Duyu Tang, Bing Qin*, Ting Liu
Harbin Institute of Technology, Harbin, China
{dytang, qinb, tliu}@ir.hit.edu.cn

- Task: Document-level sentiment analysis
- Datasets:
- IMDB (official split) & Yelp Dataset Challenge (4:1:1)

Corpus	#docs	#s/d	#w/d	$\|V\|$	#class	Class Distribution
Yelp 2013	335,018	8.90	151.6	211,245	5	.09/.09/.14/.33/.36
Yelp 2014	1,125,457	9.22	156.9	476,191	5	.10/.09/.15/.30/.36
Yelp 2015	1,569,264	8.97	151.9	612,636	5	.10/.09/.14/.30/.37
IMDB	348,415	14.02	325.6	115,831	10	.07/.04/.05/.05/.08/.11/.15/.17/.12/.18

Architecture



- Sentence-level: CNN or LSTM
- Document-level: GRU + Avg Pooling

Comparing with Other Methods

	Yelp 2013		Yelp 2014		Yelp 2015		IMDB	
	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE
Majority	0.356	3.06	0.361	3.28	0.369	3.30	0.179	17.46
SVM + Unigrams	0.589	0.79	0.600	0.78	0.611	0.75	0.399	4.23
SVM + Bigrams	0.576	0.75	0.616	0.65	0.624	0.63	0.409	3.74
SVM + TextFeatures	0.598	0.68	0.618	0.63	0.624	0.60	0.405	3.56
SVM + AverageSG	0.543	1.11	0.557	1.08	0.568	1.04	0.319	5.57
SVM + SSWE	0.535	1.12	0.543	1.13	0.554	1.11	0.262	9.16
JMARS	N/A	–	N/A	–	N/A	–	N/A	4.97
Paragraph Vector	0.577	0.86	0.592	0.70	0.605	0.61	0.341	4.69
Convolutional NN	0.597	0.76	0.610	0.68	0.615	0.68	0.376	3.30
Conv-GRNN	0.637	0.56	0.655	0.51	0.660	0.50	0.425	**2.71**
LSTM-GRNN	**0.651**	**0.50**	**0.671**	**0.48**	**0.676**	**0.49**	**0.453**	3.00

Model Analysis

	Yelp 2013		Yelp 2014		Yelp 2015		IMDB	
	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE	Accuracy	MSE
Average	0.598	0.65	0.605	0.75	0.614	0.67	0.366	3.91
Recurrent	0.377	1.37	0.306	1.75	0.383	1.67	0.176	12.29
Recurrent Avg	0.582	0.69	0.591	0.70	0.597	0.74	0.344	3.71
Bi Recurrent Avg	0.587	0.73	0.597	0.73	0.577	0.82	0.372	3.32
GatedNN	0.636	0.58	0.656	0.52	0.651	0.51	**0.430**	2.95
GatedNN Avg	0.635	0.57	**0.659**	0.52	0.657	0.56	0.416	2.78
Bi GatedNN Avg	**0.637**	**0.56**	0.655	**0.51**	**0.660**	**0.50**	0.425	**2.71**

Remarks

- Weird to use LSTM and GRU differently
  - for sentence-level and document level modeling

- Consensus: LSTM ~ GRU

Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition

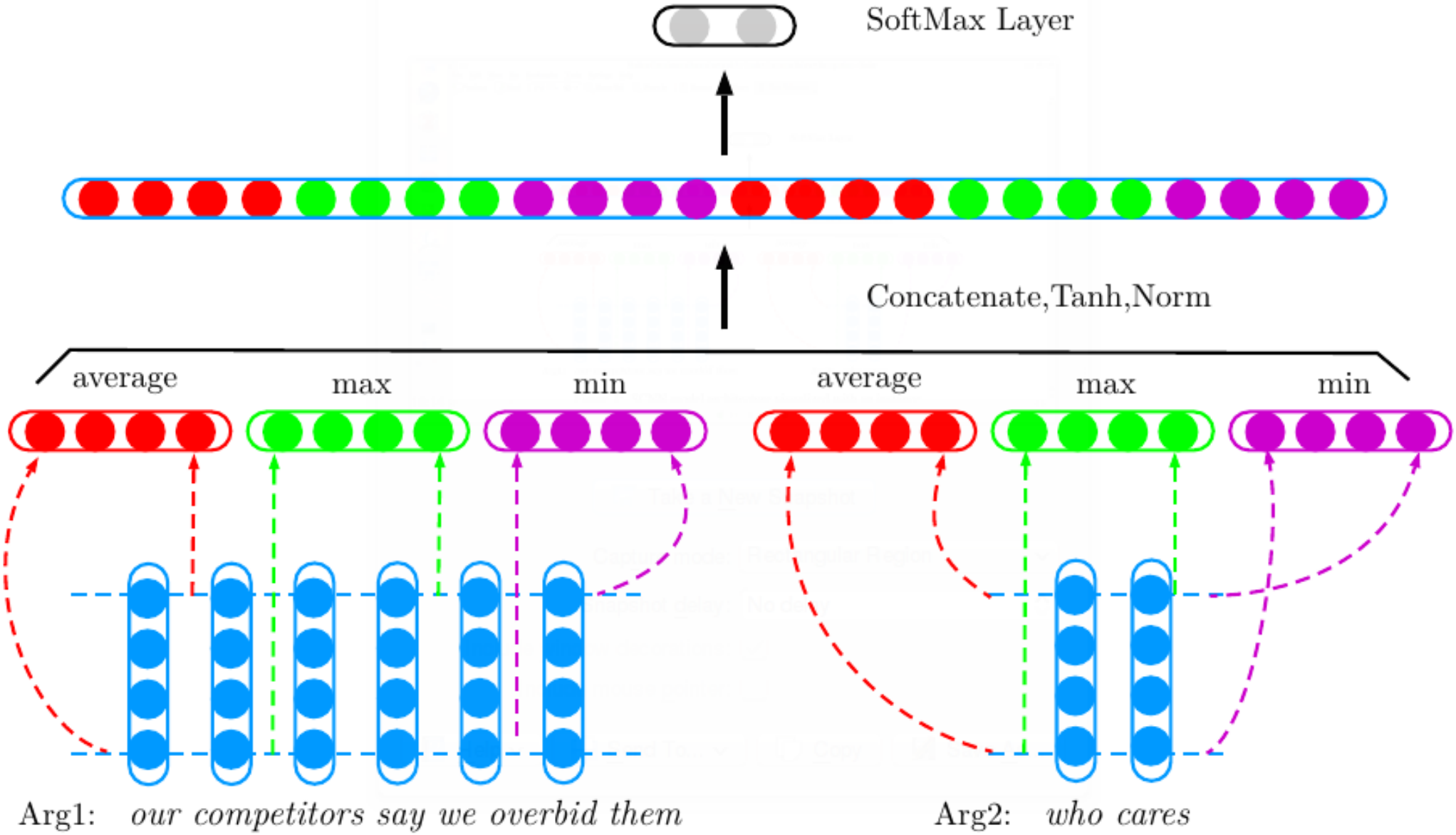
Biao Zhang[1], **Jinsong Su**[1]*, **Deyi Xiong**[2], **Yaojie Lu**[1], **Hong Duan**[1] and **Junfeng Yao**[1]
Xiamen University, Xiamen, China 361005[1]
Soochow University, Suzhou, China 215006[2]
{zb, lyj}@stu.xmu.edu.cn, {jssu, hduan, yao0010}@xmu.edu.cn
dyxiong@suda.edu.cn

- Task: To classify the relation between 2 sentences
- (successive in a paragraph, possibly)

SHORT

Arg1: Our competitions say we overbid
Arg2: Who cars
Label: COMPARISON

- Other labels: TEMPORAL , CONTINGENCY, EXPANSION


SoftMax Layer
Concatenate,Tanh,Norm
average
max
min
average
max
min
Arg1: our competitors say we overbid them
Arg2: who cares

Statistics

Relation	Positive/Negative Sentences		
	Train	Dev	Test
COMP.	1942/1942	197/986	152/894
CONT.	3342/3342	295/888	279/767
EXP.	7004/7004	671/512	574/472
TEMP.	760/760	64/1119	85/961

Results

Relation	Model	Precision	Recall	Accuracy	MacroF1
COMP. vs Other	**SVM**	22.22	60.53	63.48	32.51
	TSVM	20.53	66.45	57.74	31.37
	Add-Bro	22.79	64.47	63.10	33.68
	No-Cro	22.89	67.76	62.14	**34.22**
	RAE	18.38	62.50	54.21	28.40
	SCNN-No-Norm	21.07	54.61	63.67	30.40
	SCNN	22.00	67.76	60.42	33.22
CONT. vs Other	**SVM**	39.70	67.03	64.05	49.87
	TSVM	38.72	67.03	62.91	49.08
	Add-Bro	39.14	72.40	62.62	50.82
	No-Cro	39.50	74.19	62.81	51.56
	RAE	37.55	68.10	61.28	48.41
	SCNN-No-Norm	39.02	71.33	62.62	50.44
	SCNN	39.80	75.29	63.00	**52.04**
EXP. vs Other	**SVM**	66.35	60.10	61.38	63.07
	TSVM	66.48	61.15	61.76	63.70
	Add-Bro	65.89	58.89	60.71	62.19
	No-Cro	66.73	61.15	61.95	63.82
	RAE	58.24	70.29	56.02	63.67
	SCNN-No-Norm	59.39	74.39	58.03	66.05
	SCNN	56.29	91.11	56.30	**69.59**
TEMP. vs Other	**SVM**	15.76	68.24	67.78	25.61
	TSVM	16.26	77.65	65.68	26.88
	Add-Bro	15.10	68.24	66.25	24.73
	No-Cro	13.89	64.71	64.53	22.87
	RAE	10.02	60.00	52.96	17.17
	SCNN-No-Norm	18.26	67.06	72.94	28.71
	SCNN	20.22	62.35	76.95	**30.54**

Remarks

- Related topic: sentence pair modeling
- Related task: paragraph detection
- New dataset:
  A large annotated corpus for learning natural language inference, EMNLP, best data set or resource paper

Discourse Element Identification in Student Essays based on Global and Local Cohesion

SHORT

Wei Song[†], **Ruiji Fu**[‡], **Lizhen Liu**[†], **Ting Liu**[§]
[†]Information Engineering, Capital Normal University, Bejing 100048, China
[‡]Iflytek Research Beijing, Beijing 100083, China
[§]Harbin Institute of Technology, Harbin 150001, China
{wsong, lzliu}@cnu.edu.cn, rjfu@iflytek.com, tliu@ir.hit.edu.cn

Task: To classify the role of a sentence in an essay

Dataset: High school essays, annotated by two volunteers
NOT publicly available

Element	Definition
Introduction (I)	introduces the background and/or grabs readers' attention
Prompt (P)	restates or summarize the prompt
Thesis (T)	states the author's main claim on the issue for which he/she is arguing
Main idea (M)	asserts foundational ideas or aspects that are related to the thesis
Supporting idea (S)	provides evidence to explain or support the thesis and main ideas
Conclusion (C)	concludes the whole essay or one of the main ideas
Other (O)	doesn't fit into the above elements or makes no meaningful contribution

Table 1: Definitions of discourse elements.

Approach

- Individual classifier for each sentence
- Structure learning: Linear CRF

Local features

Position features The relative position of its paragraph (first, last or body) in the essay and its relative position (first, last or body) in the paragraph are modeled as a set of binary features. The index of the sentence is also used as a feature.

Indicator features Cue words/phrases like "我认为(in my opinion)" and "总之(in conclusion)" are used as indicators. Partial indicators are adapted from the ones used by Persing et al. (2010). More Chinese specific indicators are then augmented manually. We use a binary feature denoting a reference to the first person ("我(I)","我们(We)") in the sentence. We also use a binary feature to indicate whether the sentence contains a modal verb like "应该(should)" and "希望(hope)".

Lexical features Binary features are modeled for all connectives and adverbs, which are identified based on POS tags.

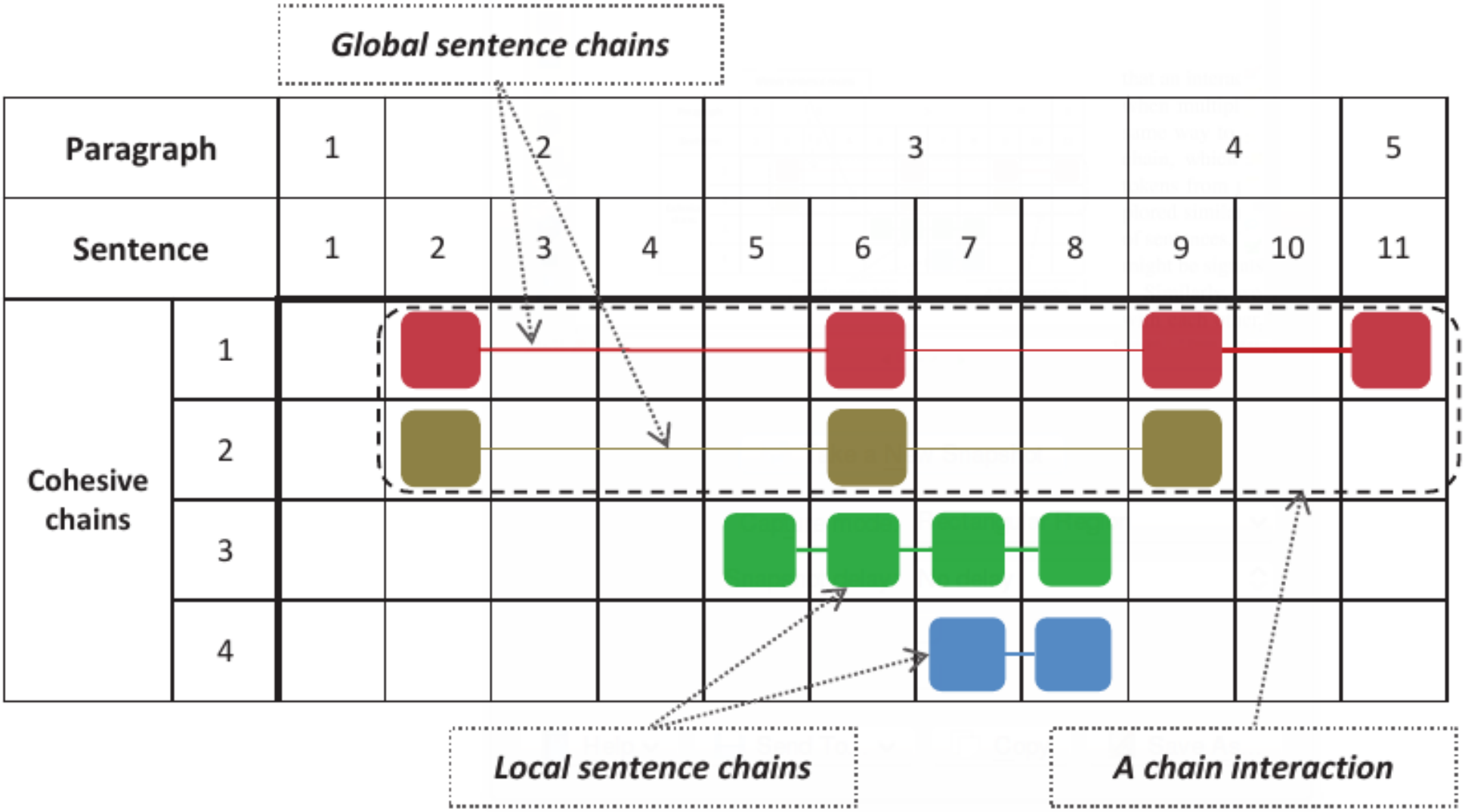
Topic and prompt features For each sentence, the cosine similarities to the essay title and to the prompt are used as features.

Cohesive Features

- Define cohesive chains
  - Identity chain: NER person, third-person resolution
  - Lexical chain:

    word2vec → cluster by threshold →

    add a link chain between sentences that contain words in a cluster
- Distinguish two cohesive chains
  - Local chain: A chain connecting <=2 paragraphs
  - Global chain: >= 3 paragraphs
- Features: # of chains
  - global-identity, global lexical, global lexical, local lexical


Global sentence chains
Paragraph
1
2
3
4
5
Sentence
1
2
3
4
5
6
7
8
9
10
11
Cohesive chains
1
2
3
4
Local sentence chains
A chain interaction

More Heuristic Features

- Global-title: (binary)
  - A sentence in a global chain AND Overlap the title?
- Chain interaction (2 binary features)
  - Two chains containing multiple sentences
  - Distinguish between GLOBAL and LOCAL interaction
- Strength features (for a sentence)
  - the number chains, the maximum and average number of covered sentences and paragraphs over chains

Experiments

Element	Features	C1			C2			C3			$avg.\triangle(F1)$
		P	R	$F1$	P	R	$F1$	P	R	$F1$	
Introduction	Basic	84.5	89.6	86.8	82.2	80.7	81.5	80.6	90.1	85.0	+3.7
	+ Cohesion	87.2	90.8	88.8	85.6	84.8	85.2	87.3	94.4	90.6	
Prompt	Basic	89.7	86.9	88.2	77.2	69.0	72.5	–	–	–	+1.9
	+ Cohesion	91.1	89.2	90.1	82.0	69.1	74.4	–	–	–	
Thesis	Basic	76.5	69.0	72.4	69.9	61.1	64.9	73.3	57.5	64.0	+5.1
	+ Cohesion	78.3	73.1	75.5	75.4	63.8	68.6	77.3	68.9	72.7	
Main idea	Basic	71.4	59.1	64.5	69.0	60.9	64.6	69.4	54.0	60.7	+5.4
	+ Cohesion	75.7	65.3	70.0	73.6	61.3	66.8	75.7	64.3	69.4	
Supporting idea	Basic	86.1	91.4	88.6	83.8	89.6	86.6	83.8	90.5	87.0	+1.8
	+ Cohesion	88.0	92.3	90.1	84.2	91.6	87.7	87.7	92.2	89.9	
Conclusion	Basic	87.2	89.9	88.4	85.6	88.5	87.0	88.1	91.0	89.5	+2.2
	+ Cohesion	89.1	91.9	90.4	86.0	90.7	88.2	92.1	94.0	93.1	

- CRF > SVM
- What is the baseline? SVM?

Model Analysis

- How is AUC computed?
- To compute AUC, we need a hyperparmeter to balance between P and R

Cohesion Feature	AUC
Global-lexical	0.712
Avg.#paras	0.670
Global-title	0.664
Max.#para	0.659
Global-interaction	0.654
Max.#sents	0.636
Avg.#sents	0.613
#Chains	0.601
Local-title	0.522
Global-identity	0.510
Local-identity	0.481
Local-interaction	0.476
Local-lexical	0.431

LONG

Comparing Word Representations for Implicit Discourse Relation Classification

Chloé Braud
ALPAGE, Univ Paris Diderot
& INRIA Paris-Rocquencourt
75013 Paris - France
chloe.braud@inria.fr

Pascal Denis
MAGNET, INRIA Lille Nord-Europe
59650 Villeneuve d'Ascq - France
pascal.denis@inria.fr

- Dataset: Penn Discourse Treebank
- Main finding:

  Dense vec > sparse

- What special in DRR?

Relation	Train	Dev	Test
Temporal	665	93	68
Contingency	3,281	628	276
Comparison	1,894	401	146
Expansion	6,792	1,253	556
Total	12,632	2,375	1,046

	All words				Head words only			
Representation	*Temp.*	*Cont.*	*Compa.*	*Exp.*	*Temp.*	*Cont.*	*Compa.*	*Exp.*
One-hot $\otimes$	21.14	50.36	34.80	59.43	11.96	43.24	17.30	**69.21**
One-hot $\oplus$	23.04	51.31	34.06	58.96	23.01	49.40	29.23	59.08
Brown 3,200 $\otimes$	20.38	50.95	34.85	61.23	11.98	43.77	16.75	68.76
Best *Brown* $\otimes$	15.52	**53.85****	30.90	61.87	22.91	45.74	25.83	68.76
Best *Brown* $\oplus$	**27.96****	49.48	31.19	**67.42****	21.84	47.36	27.52	61.38
Best *Embed.* $\otimes$	22.97	52.76**	**34.99**	61.87	**23.88**	**51.29**	**30.59**	58.59
Best *Embed.* $\oplus$	25.98*	52.50	33.15	60.17	22.48	47.48	29.82	57.45

	Temporal	*Contingency*	*Comparison*	*Expansion*
System	F1	F1	F1	F1
(Ji and Eisenstein, 2014)	26.91	51.39	35.84	**79.91**
(Rutherford and Xue, 2014)	28.69	54.42	**39.70**	70.23
repr. (Rutherford and Xue, 2014) NB	28.05	52.95	37.38	70.23
repr. (Rutherford and Xue, 2014) ME	24.79	53.39	36.46	50.00
One-hot $\otimes$ all tokens + add. features	23.26	54.41	34.34	62.57
Best all tokens only	27.96	53.85	34.99	67.42
Best all tokens + add. features	**29.30**	**55.76**	36.36	61.76

The Overall Markedness of Discourse Relations

SHORT

Lifeng Jin and Marie-Catherine de Marneffe
Department of Linguistics
The Ohio State University
{jin, mcdm}@ling.osu.edu

(Computational psycholinguistics track)

- Continuous relation:
  - E.g., causal, temporal succession, topic succession
- Discontinuous relation:
  - E.g., contradiction

- I was tired, so I drank a cup of coffee. (continuous)
- I drank a cup of coffee but I was still tired. (discontinuous)

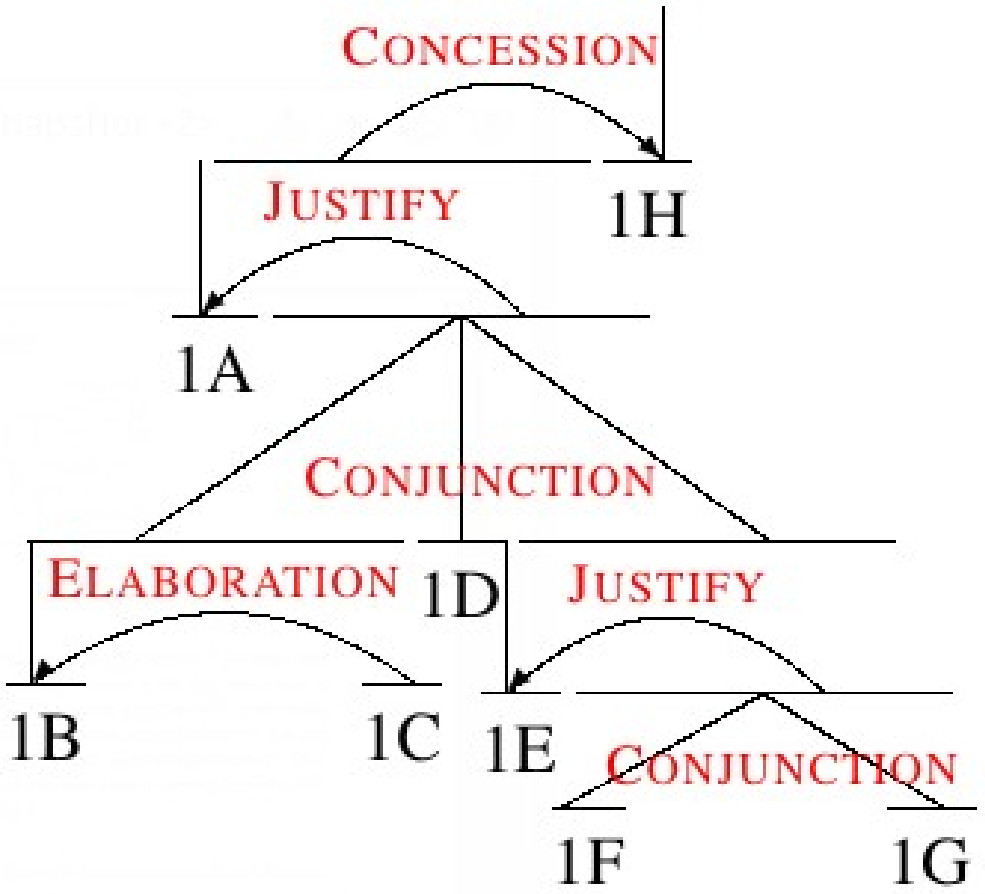
- Continuity hypothesis:
  - Sentences connected by continuous relations are easier to understand than ones connected by discontinuous relations
- The goal of the paper:
  - To propose a measure (markedness) to fit the continuity hypothesis

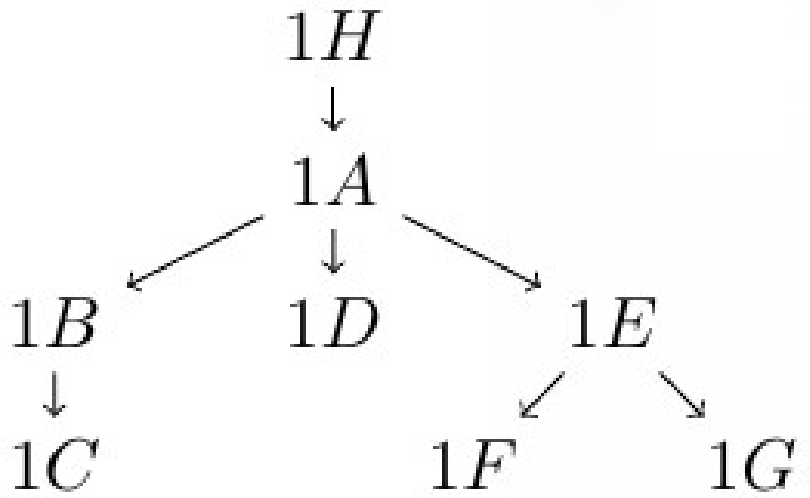
Better Document-level Sentiment Analysis from RST Discourse Parsing*

SHORT

Parminder Bhatia and **Yangfeng Ji** and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30308
parminder.bhatia243@gmail.com, jiyfeng@gatech.edu, jacobe@gatech.edu

- Re-weight the contribution of each discourse unit, based on its poistion in a dependency-like representation of the discourse structure
- Recursively propagate sentiment up through the RST parse (like RNN)





[It could have been a great movie]^{1A} [It does have beautiful scenery,]^{1B} [some of the best since Lord of the Rings.]^{1C} [The acting is well done,]^{1D} [and I really liked the son of the leader of the Samurai.]^{1E} [He was a likable chap,]^{1F} [and I hated to see him die.]^{1G} [But, other than all that, this movie is nothing more than hidden rip-offs.]^{1H}

Dataset

- Pang and Lee (2004) ~2000 reviews
- Socher et al. (2013) ~50,000 reviews

  ???

Discourse depth reweighting

$$\lambda_i = \max(0.5, 1 - d_i/6)$$

- lambda_i: coefficient; d_i: depth
- The overall prediction

$$\Psi = \sum_i \lambda_i(\boldsymbol{\theta}^\top \boldsymbol{w}_i) = \boldsymbol{\theta}^\top(\sum_i \lambda_i \boldsymbol{w}_i)$$

- w_i: BoW vector; theta=1 if postive, -1, if negative

Rhetorical Recursive Neural Network

$$\Psi_i = \tanh(K_n^{(r_i)} \Psi_{n(i)} + K_s^{(r_i)} \Psi_{s(i)})$$

- Overall document representation

  Long propagation path also addressed in the paper

$$\Psi_{\text{doc}} = \gamma \boldsymbol{\theta}^\top (\sum_i \boldsymbol{w}_i) + \Psi_{\text{rst-root}}$$

Results

	Pang & Lee	Socher et al.
Baselines		
B1. Lexicon	68.3	74.9
B2. Classifier	82.4	81.5
Discourse depth weighting		
D1. Lexicon	72.6	78.9
D2. Classifier	82.9	82.0
Rhetorical recursive neural network		
R1. No relations	83.4	85.5
R2. With relations	84.1	85.6

- Lesson: Combining the idea of RNN (or TBCNN) with traditional surface features

One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations

TACL

Yangfeng Ji and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
{jiyfeng, jacobe}@gatech.edu

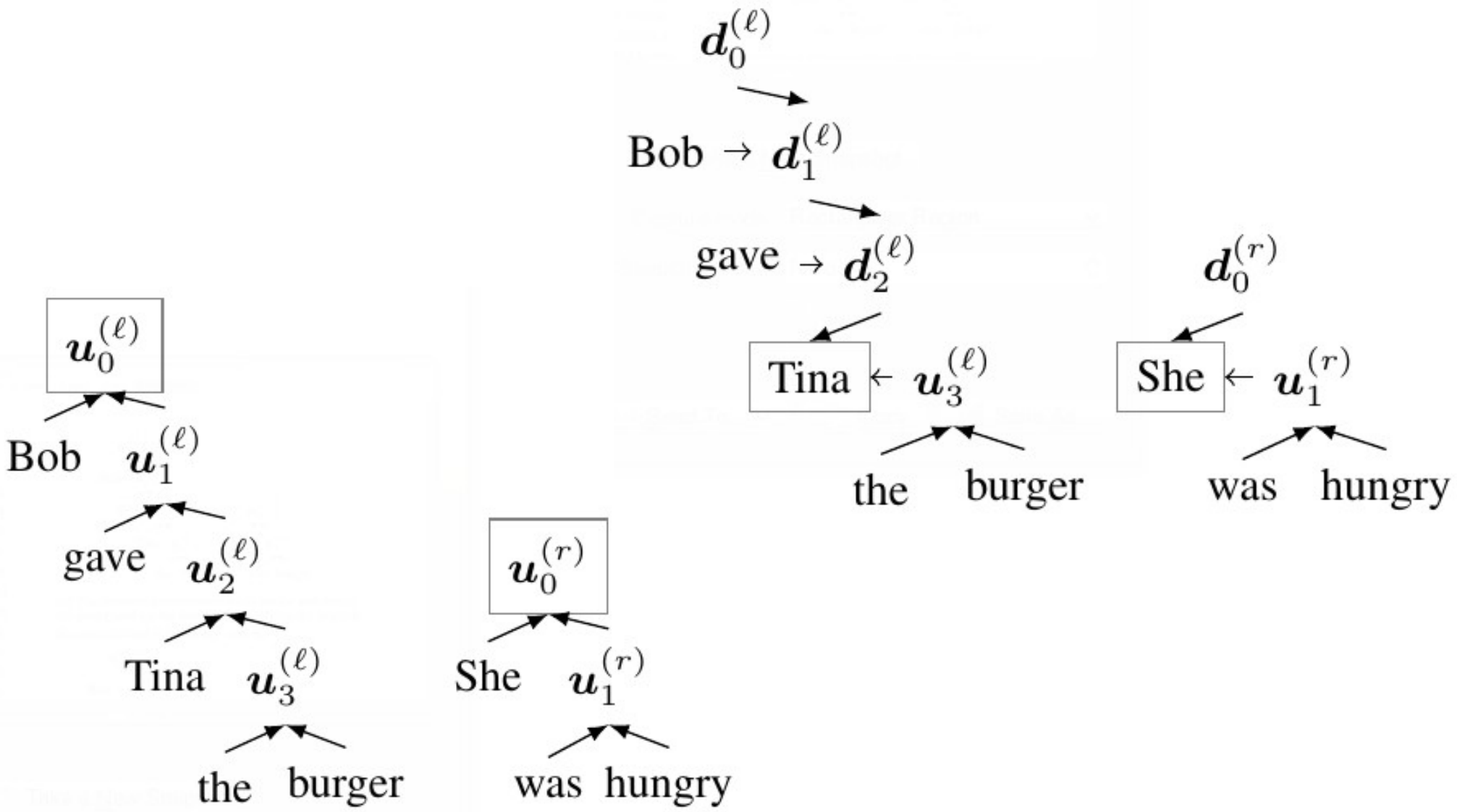
- Bob gave Tina the burger.

  She was hungry.

- Bob gave Tina the burger.

  He was hungry.

Architecture



- Decision function (scoring function)

$$\psi(y) = (\boldsymbol{u}_0^{(m)})^\top \mathbf{A}_y \boldsymbol{u}_0^{(n)} + \sum_{i,j \in \mathcal{A}(m,n)} (\boldsymbol{d}_i^{(m)})^\top \mathbf{B}_y \boldsymbol{d}_j^{(n)} + \boldsymbol{\beta}_y^\top \boldsymbol{\phi}_{(m,n)} + b_y, \tag{5}$$

  - A(m, n): Aligned entity mentions (Why does it matter?)
  - Phi: Surface features
  - A, B: Low rank approximation

$$\mathbf{A}_y = \boldsymbol{a}_{y,1} \boldsymbol{a}_{y,2}^\top + \mathrm{diag}(\boldsymbol{a}_{y,3})$$

- Cost function: Hinge loss

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{y' : y' \neq y^*} \max\left(0, 1 - \psi(y^*) + \psi(y')\right) + \lambda ||\boldsymbol{\theta}||_2^2$$

Closing the Gap:
Domain Adaptation from Explicit to Implicit Discourse Relations

SHORT

Yangfeng Ji **Gongbo Zhang** **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
{jiyfeng,gzhang64,jacobe}@gatech.edu

- Explicit relation v.s. Implicit relation
- Weak supervision: Use explicit data to train models for implicit data
- Poor performance: <== linguistically dissimilar (?)
- The goal of this paper: Domain adaption
  - Feature representation learning + Resampling

Learning robust features by Denoising AE

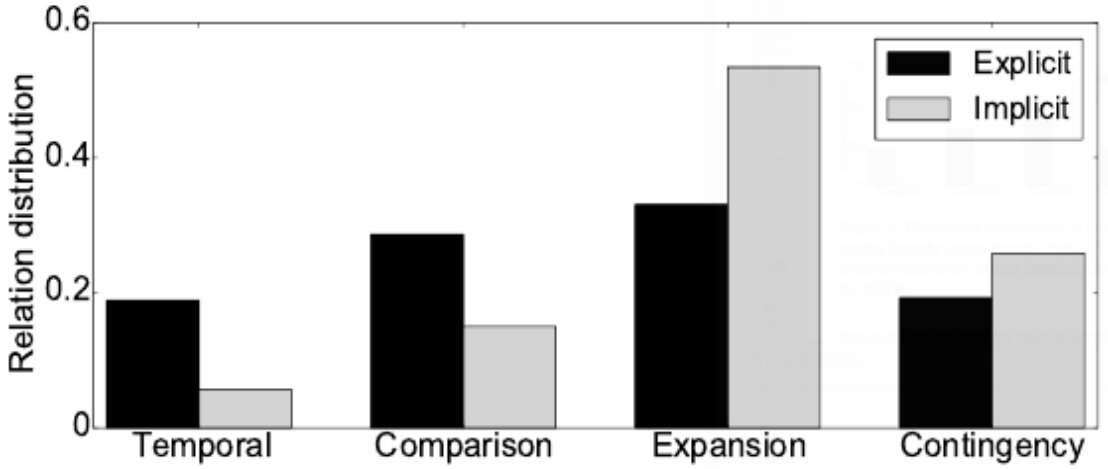
$$\min_{\mathbf{W}} E_{\tilde{\boldsymbol{x}}_i|\boldsymbol{x}_i}[\|\boldsymbol{x}_i - \mathbf{W}\tilde{\boldsymbol{x}}_i\|^2]$$

- x_tilde: corrupted features
  - Gaussian noise for continuous features
  - Drop out for binary features
- x: features, ~1e5 dimensions
- W: prohibitively large
  - Trick: kappa pivot features (Blitzer et al., 2006)

- Features (proposed in previous work)
  - Lexical
  - Syntactic
  - Others

Resampling with minimal supervision

- Matching the distribution



- Instance weighting
  - Require sampled instance having at least tau-cosine similarity with at least one sample in the target domain

Results

Surface Features	+Rep. Learning	+Resampling	Relations				Average F_1
			TEMP.	COMP.	EXP.	CONT.	
Implicit → Implicit							
1. FULL			24.15	28.87	68.84	43.45	41.32
Explicit [PDTB] → Implicit							
2. FULL	No	No	17.13	20.54	50.55	36.14	31.04
3. FULL	No	Yes	15.38	23.88	62.04	35.29	34.14
4. FULL	Yes	No	17.53	22.77	50.85	36.43	31.90
5. FULL	Yes	Yes	17.05	22.00	63.51	38.23	35.20
6. PIVOT	No	No	17.33	23.89	53.53	36.22	32.74
7. PIVOT	No	Yes	17.73	25.39	62.65	36.02	35.44
8. PIVOT	Yes	No	18.66	25.86	63.37	38.87	36.69
9. PIVOT	Yes	Yes	19.26	25.74	68.08	41.39	38.62
Explicit [PDTB + CNN] → Implicit							
10. PIVOT	Yes	Yes	20.35	26.32	68.92	42.25	39.46

Discourse Planning with an N-gram Model of Relations

SHORT

Or Biran
Columbia University
orb@cs.columbia.edu

Kathleen McKeown
Columbia University
kathy@cs.columbia.edu

- Generate a comparison story based on an ontology
  - Pattern: “the [predicate(s)] of [subject] (is/are) [object(s)]”
- Main idea: using discourse relations to improve discourse planning
- Evaluation: Crowd-sourced human evalution

Message (Multi-)graph

- Vertex: A message
- Edge: A potential relation

  (annotated according to predicates)

- Goal: To hind a Hamiltonian path through the selected subgraph

N-gram model on relations

$$P(r_i|r_{i-n},\ldots,r_{i-1}) = \frac{C(r_{i-n},\ldots,r_{i-1},r_i)}{C(r_{i-n},\ldots,r_{i-1})}$$

- Choosing the order: 4 messages in total---tractable for even brute force search.

The birth place of Allen J. Ellender is Montegut, Louisiana, while the death place of Allen J. Ellender is Maryland. The birth place of Robert E. Quinn is Phoenix, Rhode Island. Subsequently, the death place of Robert E. Quinn is Rhode Island.

The birth place of Allen J. Ellender is Montegut, Louisiana. In comparison, the birth place of Robert E. Quinn is Phoenix, Rhode Island. The death place of Robert E. Quinn is Rhode Island, but the death place of Allen J. Ellender is Maryland.

Figure 1: Sample pair of comparison stories

Results

- Base: random order
- PDTB: n-gram model on PDTB
- Wiki: n-gram model on Wikipedia, annotated by a discourse parser
- Equal: two models are the same in human eval

	Quality comparison			Avg. score	
	Base	Equal	Pdtb	Base	Pdtb
Of. Holder	27.4%	30.2%	**42.5%**	3.67	**3.76**
TV Show	34.3%	25.7%	**40%**	3.79	**3.8**
Mil. Unit	32.3%	23.2%	**44.4%**	3.69	**3.84**
River	**39.2%**	23.5%	37.3%	3.71	**3.72**
Total	34%	25%	**41%**	3.72	**3.78**

Table 1: Results for the comparison between the PDTB n-gram model and the baseline

	Quality comparison			Avg. score	
	Pdtb	Equal	Wiki	Pdtb	Wiki
Of. Holder	33.6%	14.5%	**51.8%**	3.51	**3.65**
TV Show	43.2%	8.1%	**48.6%**	3.62	**3.65**
Mil. Unit	40.4%	14.4%	**45.2%**	3.65	**3.67**
River	41.1%	11.2%	**47.7%**	3.68	**3.7**
Total	39.6%	12%	**48.4%**	3.61	**3.67**

A Hierarchical Neural Autoencoder for Paragraphs and Documents

ACL-LONG

Jiwei Li, Minh-Thang Luong and Dan Jurafsky
Computer Science Department, Stanford University, Stanford, CA 94305, USA
jiweil, lmthang, jurafsky@stanford.edu

- Goal: To encode a sentence, and to decode it

ence. While only a first step toward generating coherent text units from neural models, our work has the potential to significantly impact natural language generation and summarization[1].

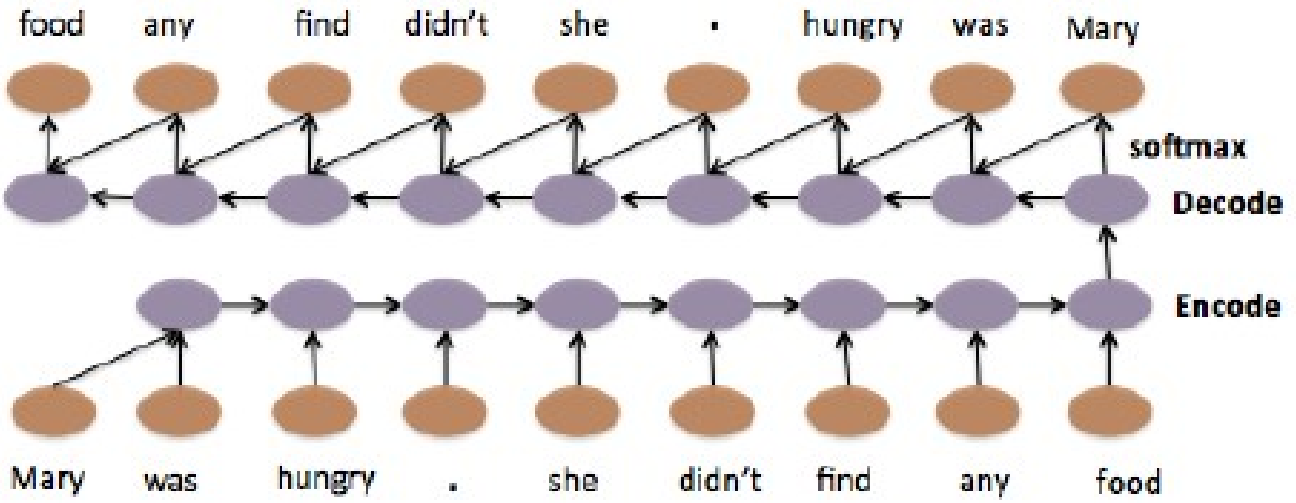


Figure 1: Standard Sequence to Sequence Model.

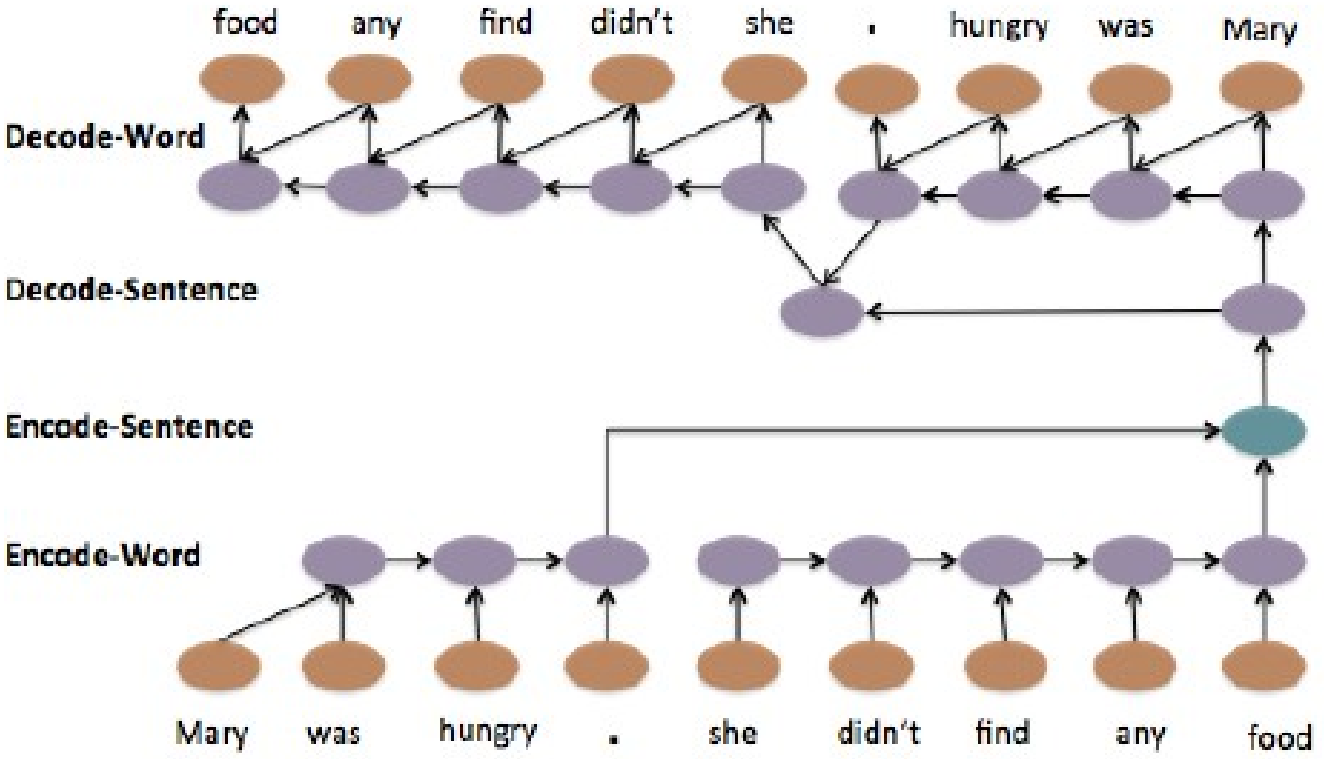
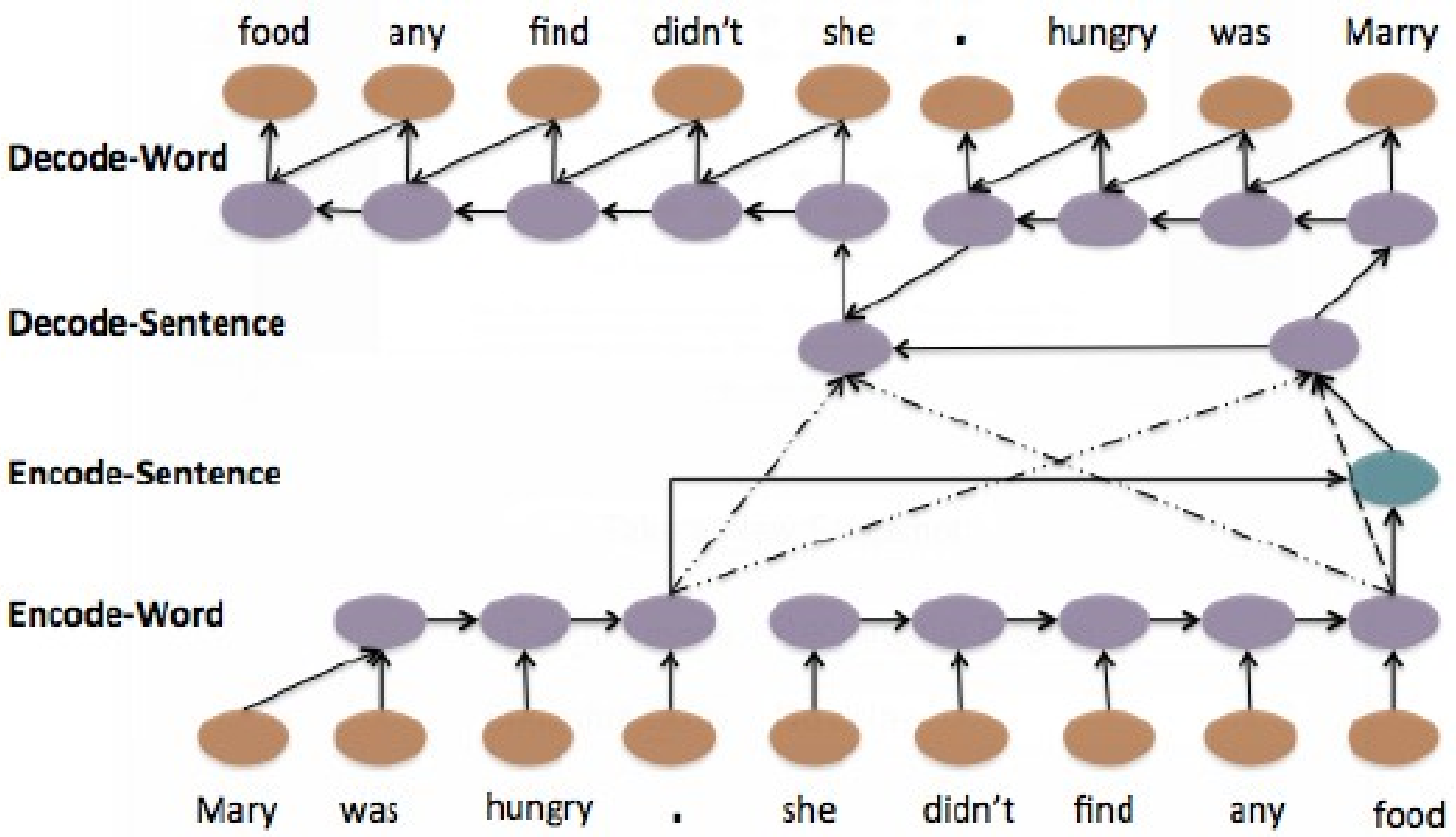


Figure 2: Hierarchical Sequence to Sequence Model.



- Evaluation
  - BLEU, ROUGE, Coherence

paris is the capital and most populous city of france . situated on the seine river , in the north of the country , it is in the centre of the le-de-france region . the city of paris has a population of 2273305 inhabitants . this makes it the fifth largest city in the european union measured by the population within the city limits .
paris is the capital and most populated city in france . located in the <unk> , in the north of the country , it is the center of <unk> . paris , the city has a population of <num> inhabitants . this makes the eu ' s population within the city limits of the fifth largest city in the measurement .

on every visit to nyc , the hotel beacon is the place we love to stay . so conveniently located to central park , lincoln center and great local restaurants . the rooms are lovely . beds so comfortable , a great little kitchen and new wizz bang coffee maker . the staff are so accommodating and just love walking across the street to the fairway supermarket with every imaginable goodies to eat .
every time in new york , lighthouse hotel is our favorite place to stay . very convenient , central park , lincoln center , and great restaurants . the room is wonderful , very comfortable bed , a kitchenette and a large explosion of coffee maker . the staff is so inclusive , just across the street to walk to the supermarket channel love with all kinds of what to eat .

Recursive Deep Models for Discourse Parsing

Jiwei Li[1], Rumeng Li[2] and Eduard Hovy[3]
[1]Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2]School of EECS, Peking University, Beijing 100871, P.R. China
[3]Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
jiweil@stanford.edu alicerumeng@foxmail.com ehovy@andrew.cmu.edu

- Dataset: Rhetorical Structure Theory Discourse Treebank (RST-DT)
- 385 documents, 347 for training (5-fold), 49 for testing
- Each doc represented as a tree
  - Elementary Discourse Units (EDUs): Clauses
  - Relations: hypotactic v.s. paratactic

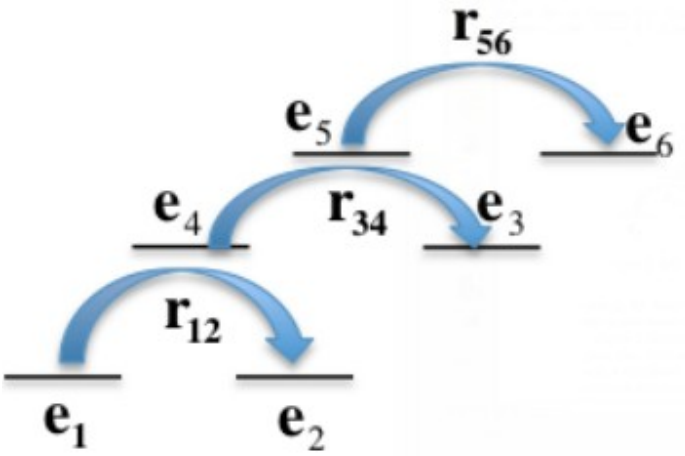
EDU Modeling

- Standard RAE

Discourse Parsing

- 2-step strategy
  - Binary classifier: To determine whether two adjacent text units should be merged to form a new subtree

$$t_{\text{binary}}(e_1, e_2) = 1, \quad t_{\text{binary}}(e_3, e_4) = 1,$$
$$t_{\text{binary}}(e_2, e_3) = 0, \quad t_{\text{binary}}(e_3, e_6) = 0,$$
$$t_{\text{binary}}(e_5, e_6) = 1$$



$$L^{\text{binary}}_{(e_i,e_j)} = f(G_{\text{binary}} * [h_{e_i}, h_{e_j}] + b_{\text{binary}})$$

$$p[t_{\text{binary}}(e_i, e_j) = 1] = g(U_{binary} \cdot L^{\text{binary}}_{(e_i,e_j)} + b^*_{\text{binary}})$$

  - Multi-class classifier: To determine which relation

Inference

- Choose the parse tree with max. prob.
- Dynamic programming, keeping 10 options at each time

Wrap Up

Discourse Analysis

- Document-level classification
  - Traditional document classification → trivial
  - Topic modeling → LDA or variants
  - Document-level sentiment analysis → New research topic
  - Shall we use discourse parse tree structures?
- Discourse relation classification
  - Sentence pair modeling
  - Paraphrase detection
- Discourse parsing (TODO)

TODO list

- Datasets
  - Penn Discourse Treebank
  - IMDB, Yelp
- Discourse parser
  - …