组会

许晓丹

20.10.31

Enhancing Neural Data-To-Text Generation Models with External Background Knowledge

Shuang Chen[1]*, **Jinpeng Wang**[2], **Xiaocheng Feng**[1], **Feng Jiang**[1,3], **Bing Qin**[1], **Chin-Yew Lin**[2]
[1] Harbin Institute of Technology, Harbin, China
[2] Microsoft Research Asia [3] Peng Cheng Laboratory
hitercs@gmail.com, {jinpwa, cyl}@microsoft.com,
{xcfeng, qinb}@ir.hit.edu.cn, fjiang@hit.edu.cn

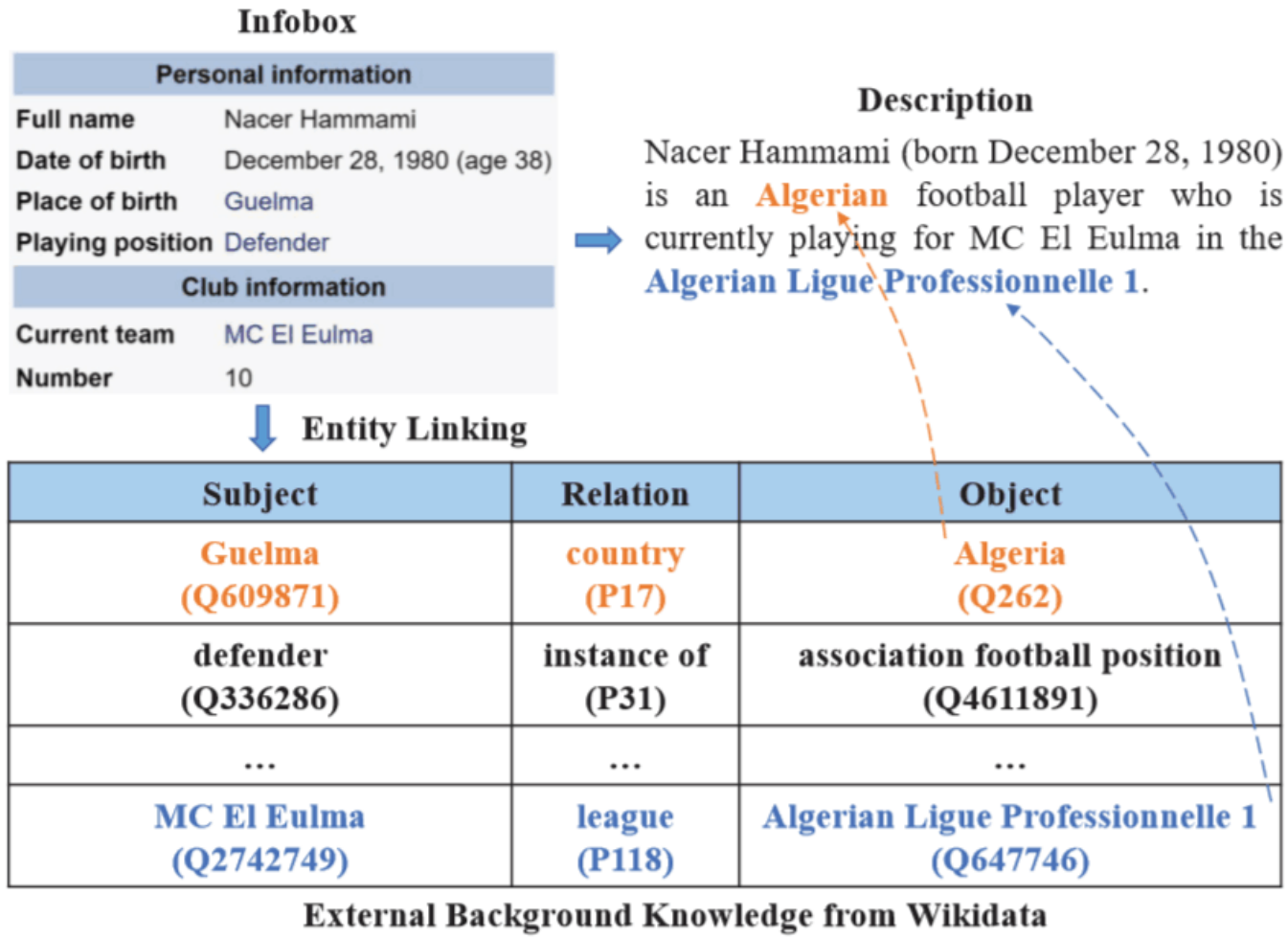


Figure 1: An example of generating description from a Wikipedia infobox. External background knowledge expanded from the infobox is helpful for generation.

Work：

- Relevant external knowledge, encoded as a temporary memory, and combines this knowledge with the context representation of data before generating words;
- Propose a dual-attention mechanism;
- KBGain.

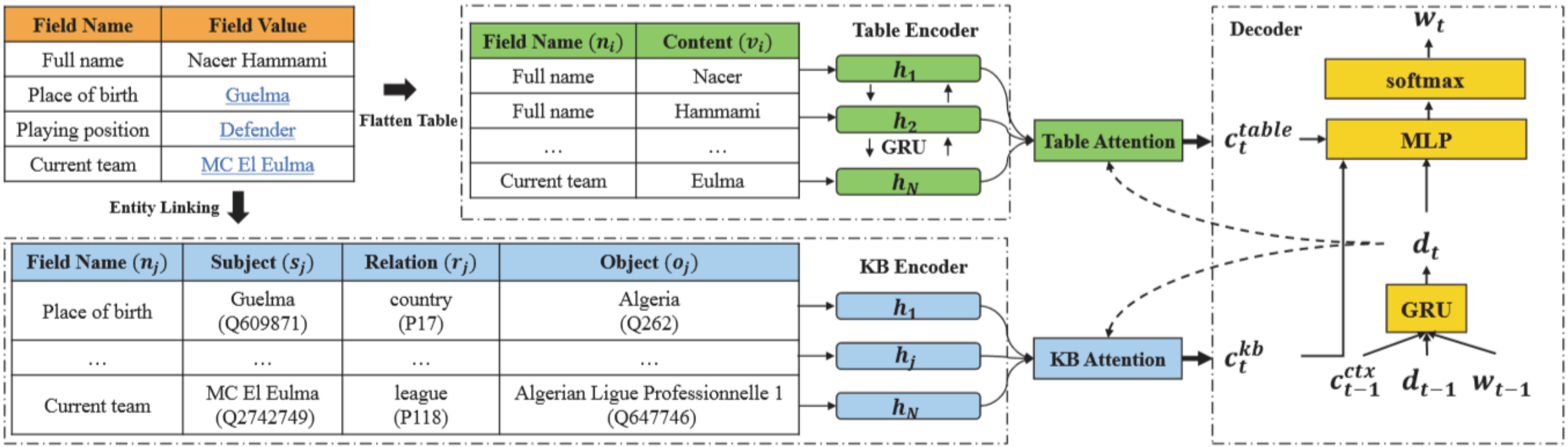


Figure 2: A diagram of the knowledge base enhanced neural data-to-text generation model. First, we transform the table into a flattened sequence, extract entities mentioned in the field value of the infobox and link them to Wikidata where we can retrieve relevant facts. Then, the table contents and external knowledge base facts are carefully encoded. Finally, a single layer GRU decoder with a dual attention mechanism decides which part of information should be used for generation.

KBGain：

- KBGain measures the portion of learnable tokens in the references co-occurred with their corresponding external KB entries but filter out those tokens which could also co-occurred with the infobox.

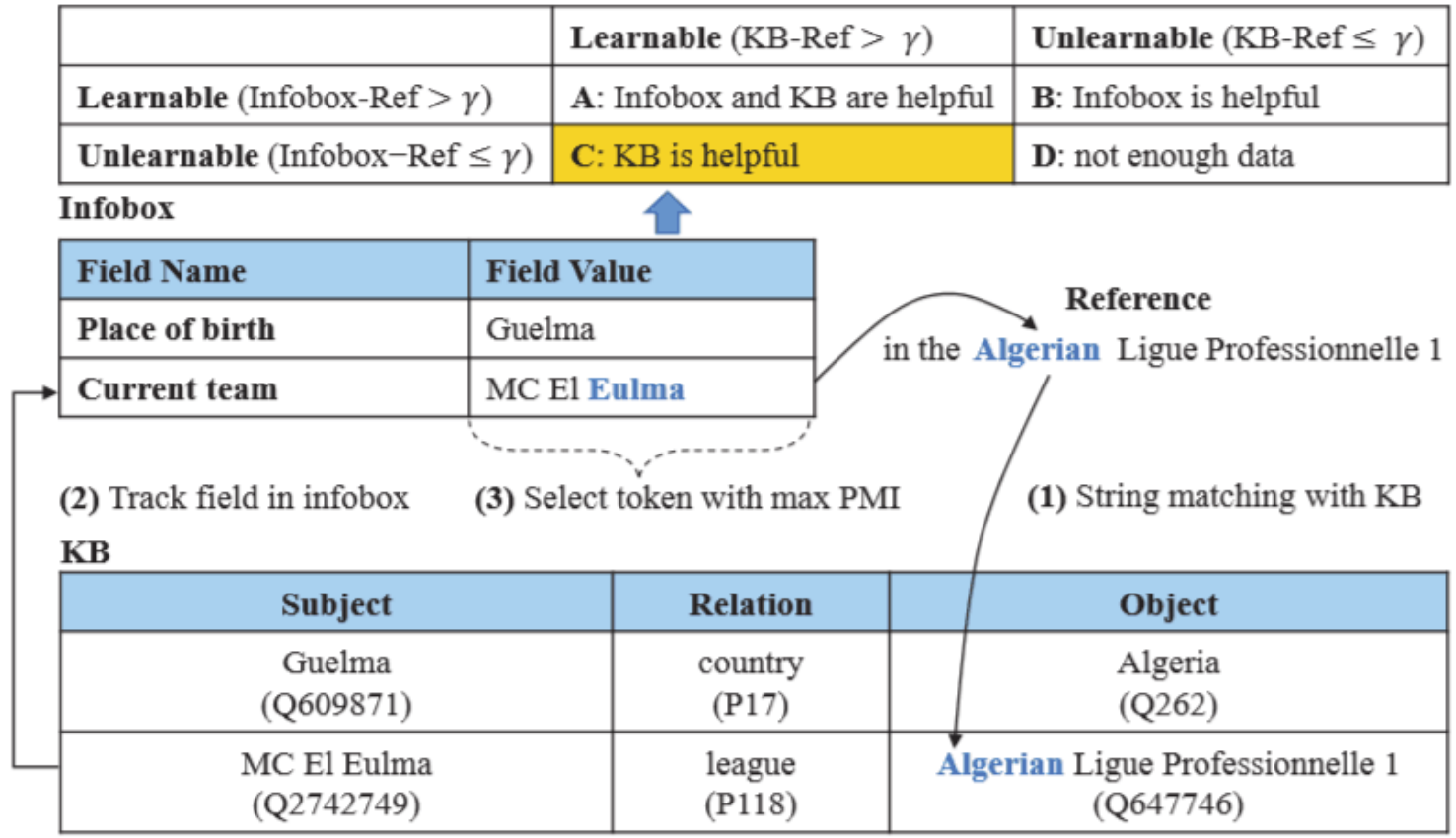


Figure 3: An illustration of KBGain metric.

Comparison:

- Baseline: Seq2seq + copy
- DataSet: WikiBio and 20 new datasets

Dataset	Seq2Seq+Copy	KBAtt	Δ
Single	43.60	47.70	4.10
Station	54.30	56.77	2.47
Australian_place	43.73	45.73	2.00
Album	41.04	42.77	1.73
NRHP	48.97	50.43	1.46
Airport	45.17	46.61	1.44
Book	36.07	37.49	1.42
Automobile	18.64	19.95	1.31
Building	24.13	25.21	1.08
UK_school	33.64	34.72	1.08
School	37.30	38.33	1.03
Football_club_season	46.05	47.02	0.97
UK_place	41.46	42.38	0.92
Military_unit	37.74	38.58	0.84
Military_conflict	18.58	19.21	0.63
WikiBio	44.28	44.59	0.31
Television_episode	73.59	73.87	0.28
NCAA_team_season	87.37	87.58	0.21
French_commune	90.14	90.33	0.19
Settlement	77.59	77.68	0.09
Video_game	29.54	29.48	-0.06

Table 2: BLEU-4 score with two generation models on 21 datasets.

Model	BLEU-4
Table NLM (Lebret et al., 2016)	34.70
Table2Seq (Bao et al., 2018)	40.26
Order Planning, full model (Sha et al., 2017)	43.91
Field-gating Seq2Seq, full model (Liu et al., 2017)	44.71
Seq2Seq+Copy	44.28
KBAtt	44.59

Table 3: BLEU-4 scores (%) on WikiBio dataset.

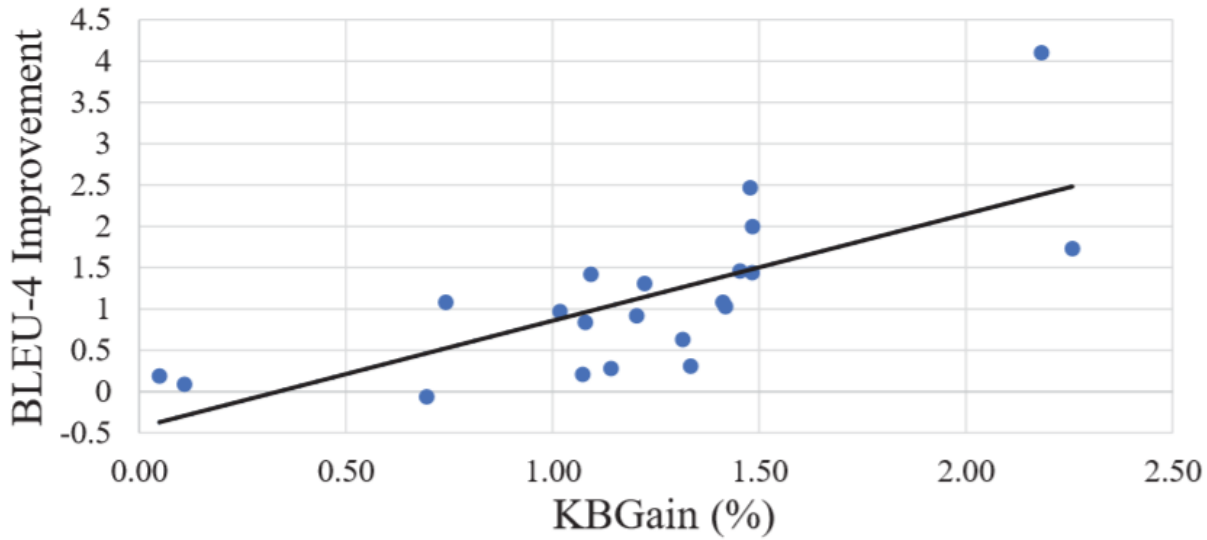
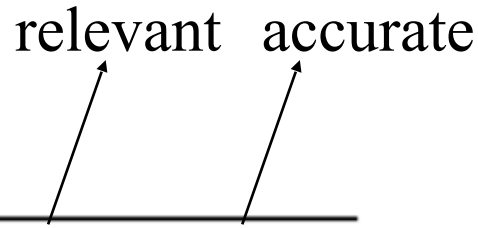


Figure 4: Strong correlation ($\rho = 0.716$) between KB-Gain (x) and absolute BLEU-4 improvement (y).



Dataset	**Model**	P_1	R_1	F_1	P_2
WikiBio	Seq2Seq+Copy	69.07%	62.45%	65.59%	88.35%
WikiBio	KBAtt	71.81%	62.45%	66.80%	91.85%
Album	Seq2Seq+Copy	73.33%	71.84%	72.58%	84.38%
Album	KBAtt	81.70%	78.37%	80.00%	90.64%

Table 4: Human based evaluation results.

Data-to-text Generation with Entity Modeling

Ratish Puduppully and **Li Dong** and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
r.puduppully@sms.ed.ac.uk li.dong@ed.ac.uk mlap@inf.ed.ac.uk

TEAM	Inn1	Inn2	Inn3	Inn4	...	R	H	E	...
Orioles	1	0	0	0	...	2	4	0	...
Royals	1	0	0	3	...	9	14	1	...

BATTER	H/V	AB	R	H	RBI	TEAM	...
C. Mullins	H	4	2	2	1	Orioles	...
J. Villar	H	4	0	0	0	Orioles	...
W. Merrifield	V	2	3	2	1	Royals	...
R. O'Hearn	V	5	1	3	4	Royals	...
...	...	...	...	...	...	...	

PITCHER	H/V	W	L	IP	H	R	ER	BB	K	...
A. Cashner	H	4	13	5.1	9	4	4	3	1	...
B. Keller	V	7	5	8.0	4	2	2	2	4	...
...	...	...	...	...	...	...	...			

Inn1: innings, R: runs, H: hits, E: errors, AB: at-bats, RBI: runs-batted-in, H/V: home or visiting, W: wins, L: losses, IP: innings pitched, ER: earned runs, BB: walks, K: strike outs.

KANSAS CITY, Mo. – **Brad Keller** kept up his recent pitching surge with another strong outing. **Keller** gave up a home run to the first batter of the game – **Cedric Mullins** – but quickly settled in to pitch eight strong innings in the Kansas City **Royals**' 9–2 win over the Baltimore **Orioles** in a matchup of the teams with the worst records in the majors. **Keller** (7–5) gave up two runs and four hits with two walks and four strikeouts to improve to 3–0 with a 2.16 ERA in his last four starts. **Ryan O'Hearn** homered among his three hits and drove in four runs, **Whit Merrifield** scored three runs, and **Hunter Dozier** and **Cam Gallagher** also went deep to help the **Royals** win for the fifth time in six games on their current homestand. With the scored tied 1–1 in the fourth, **Andrew Cashner** (4–13) gave up a sacrifice fly to **Merrifield** after loading the bases on two walks and a single. **Dozier** led off the fifth inning with a 423-foot home run to left field to make it 3-1. The **Orioles** pulled within a run in the sixth when **Mullins** led off with a double just beyond the reach of **Dozier** at third, advanced to third on a fly ball and scored on **Trey Mancini**'s sacrifice fly to the wall in right. The **Royals** answered in the bottom of the inning as **Gallagher** hit his first home run of the season...

BATTER	PITCHER	SCORER	EVENT	TEAM	INN	RUNS	...
C. Mullins	B. Keller	-	Home run	Orioles	1	1	...
H. Dozier	A. Cashner	W. Merrifield	Grounded into DP	Royals	1	1	...
W. Merrifield	A. Cashner	B. Goodwin	Sac fly	Royals	4	2	...
H. Dozier	A. Cashner	-	Home run	Royals	4	3	...
...	...	...	...	...	...	...	...

Figure 1: MLB statistics tables and game summary. The tables summarize the performance of the two teams and of individual team members who played as batters and pitchers as well as the most important events (and their actors) in each play. Recurring entities in the summary are boldfaced and colorcoded, singletons are shown in black.

Motivation：

- General neural models Treat entity as ordinary tokens;
- Descriptive texts are often characterized as “entity coherent” which means that their coherence is based on the way entities are introduced and discussed in the discourse.

Work: Propose an entity-centric neural architecture

- Entity-special representations which are dynamically updated as text is generated;
- Hierarchical attention.

DataSet

- Benchmark ROTOWIRE dataset
- A new dataset for MLB

Basic Model: Encoder-Decoder with Conditional Copy

- Encoder:

$$\{r_{j,l}\}_{l=1}^{L} \longrightarrow \mathbf{r}_j = \mathrm{ReLU}(\mathbf{W}_r[\mathbf{r}_{j,1}; \mathbf{r}_{j,2}; ...; \mathbf{r}_{j,L}] + \mathbf{b}_r)$$

$$\downarrow$$

$$\{\mathbf{e}_j\}_{j=1}^{|r|}$$

- Decoder: LSTM with copy mechanisum

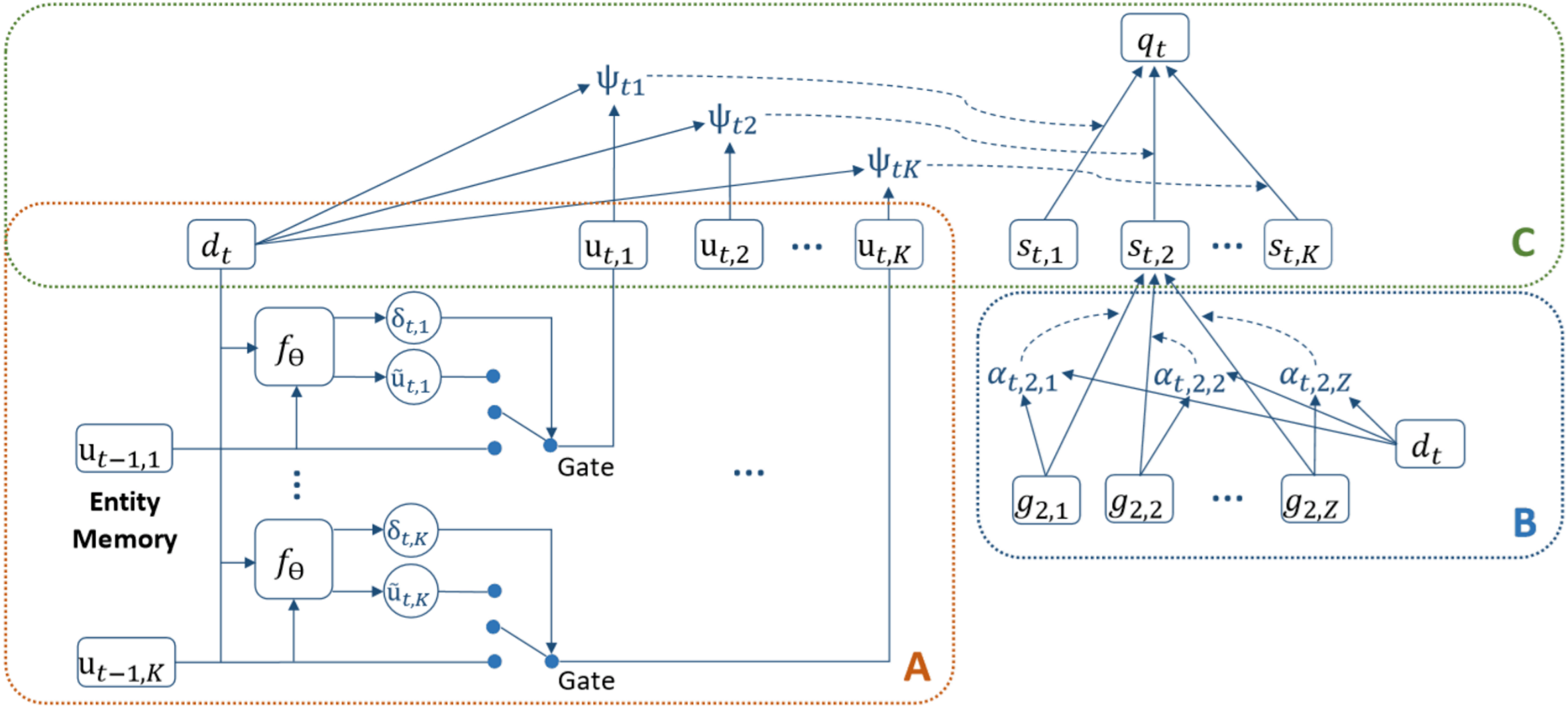
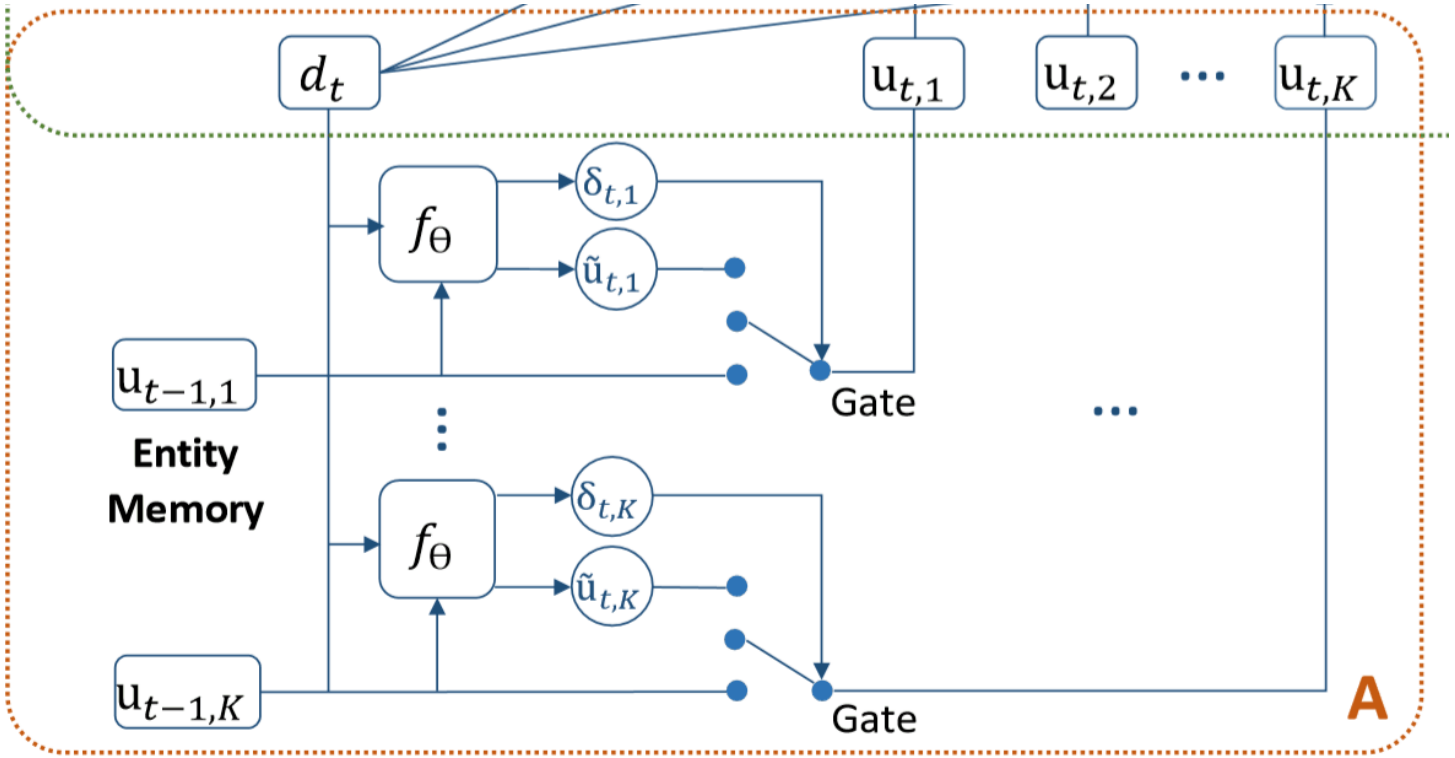


Figure 2: Diagram of entity memory network (block A) and hierarchical attention (blocks B and C). Module f_θ represents update equations (6)–(8) where θ is the set of trainable parameters. The gate represents the entity memory update (Equation (9)). Block B covers Equations (10) and (11), and block C Equations (12) and (13).

Entity Memory



$$\mathbf{x}_k = \sum_j (\mathbb{1}[r_{j,2} = k]\mathbf{r}_j) / \sum_j \mathbb{1}[r_{j,2} = k]$$

$$\mathbf{u}_{t=-1,k} = \mathbf{W}_i \mathbf{x}_k$$

$$\boldsymbol{\delta}_{t,k} = \boldsymbol{\gamma}_t \odot \sigma(\mathbf{W}_e \mathbf{d}_t + \mathbf{b}_e + \mathbf{W}_f \mathbf{u}_{t-1,k} + \mathbf{b}_f)$$

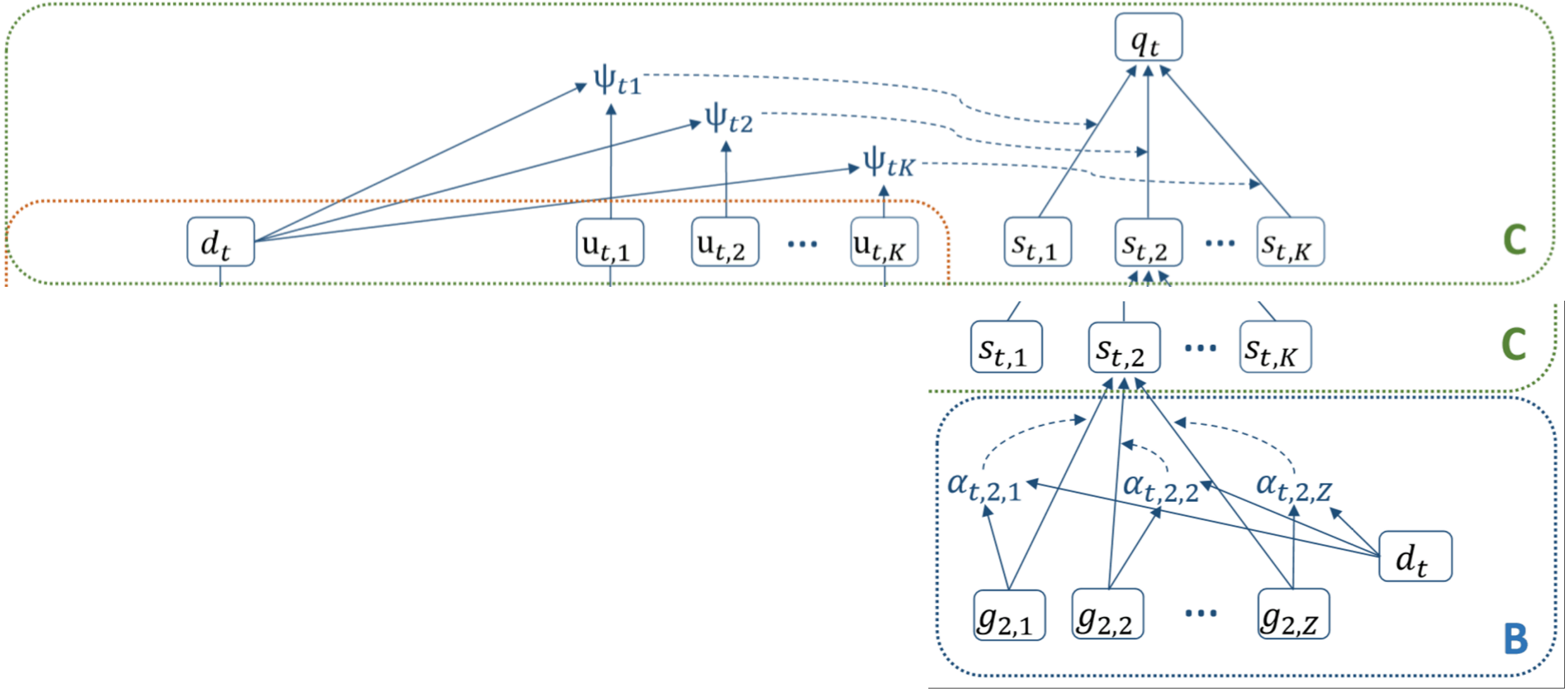
$$\mathbf{u}_{t,k} = (1 - \boldsymbol{\delta}_{t,k}) \odot \mathbf{u}_{t-1,k} + \boldsymbol{\delta}_{t,k} \odot \tilde{\mathbf{u}}_{t,k}$$

GRU

Hierarchical Attention

$\mathbf{e}_j \longrightarrow \mathbf{g}_{k,z}$

$\mathrm{j} \longleftrightarrow (\mathrm{k}, \mathrm{z})$



Comparison

	ROTOWIRE	MLB
Vocab Size	11.3K	38.9K
# Tokens	1.5M	14.3M
# Instances	4.9K	26.3K
Avg Length	337.1	542.05
# Record Types	39	53
Avg Records	628	565

Table 1: Vocabulary size, number of tokens, number of instances (i.e., record-summary pairs), average summary length, number of record types and average number of records in ROTOWIRE and MLB datasets.

RW	RG		CS		CO	BLEU
	#	P%	P%	R%	DLD%	
TEMPL	**54.23**	**99.94**	26.99	**58.16**	14.92	8.46
WS-2017	23.72	74.80	29.49	36.18	15.42	14.19
NCP+CC	34.28	87.47	34.18	51.22	18.58	**16.50**
ENT	30.11	92.69	**38.64**	48.51	**20.17**	16.12

MLB	RG		CS		CO	BLEU
	#	P%	P%	R%	DLD%	
TEMPL	**59.93**	**97.96**	22.82	**68.46**	10.64	3.81
ED+CC	18.69	92.19	**62.01**	50.12	25.44	9.69
NCP+CC	17.93	88.11	60.48	55.13	**26.71**	9.68
ENT	21.35	88.29	58.35	61.14	24.51	**11.51**

Table 2: Evaluation on ROTOWIRE (RW) and MLB test sets using relation generation (RG) count and precision, content selection (CS) precision and recall, content ordering (CO) in normalized Damerau-Levenshtein distance, and BLEU.

Ablation Study

RW	RG		CS		CO	BLEU
	#	P%	P%	R%	DLD%	
ED+CC	22.68	79.40	29.96	34.11	16.00	14.00
+Hier	30.76	93.02	33.99	44.79	19.03	14.19
+Dyn	27.93	90.85	34.19	42.27	18.47	15.40
+Gate	31.84	91.97	36.65	48.18	19.68	15.97

MLB	RG		CS		CO	BLEU
	#	P%	P%	R%	DLD%	
ED+CC	18.69	92.65	62.29	51.36	25.93	9.55
+Hier	19.02	93.71	62.84	52.12	25.72	10.38
+Dyn	20.28	89.19	58.19	58.94	24.49	10.85
+Gate	21.32	88.16	57.36	61.50	24.87	11.13

Table 3: Ablation results on ROTOWIRE (RW) and MLB development set using relation generation (RG) count and precision, content selection (CS) precision and recall, content ordering (CO) in normalized Damerau-Levenshtein distance, and BLEU.

The **Houston Rockets** (18–5) defeated the **Denver Nuggets** (10–13) 108–96 on Tuesday at the Toyota Center in Houston. The **Rockets** had a strong first half where they outscored . . . The **Rockets** were led by **Donatas Motiejunas**, who scored a game–high of 25 points . . . **James Harden** also played a factor in the win, as he went 7–for . . . Coming off the bench, **Donatas Motiejunas** had a big game and finished with 25 points . . . The only other player to reach double figures in points was **Arron Afflalo**, who came off the bench for 12 points . . . Coming off the bench, **Arron Afflalo** chipped in with 12 points . . . The **Nuggets**' next game will be on the road against the Boston Celtics on Friday, while the **Nuggets** will travel to Boston to play the Celtics on Wednesday.
The **Houston Rockets** (18–5) defeated the **Denver Nuggets** (10–13) 108–96 on Monday at the Toyota Center in Houston. The **Rockets** were the superior shooters in this game, going . . . The **Rockets** were led by the duo of **Dwight Howard** and **James Harden**. **Howard** shot 9–for–11 from the field and . . . **Harden** on the other hand recorded 24 points (7–20 FG, 2–5 3Pt, 8–9 FT), 10 rebounds and 10 assists, The only other Nugget to reach double figures in points was **Arron Afflalo**, who finished with 12 points (4–17 FG,. . . The **Rockets**' next game will be on the road against the New Orleans Pelicans on Wednesday, while the **Nuggets** will travel to Los Angeles to play the Clippers on Friday.

Table 4: Examples of model output for NCP+CC (top) and ENT (bottom) on ROTOWIRE. Recurring entities in the summaries are boldfaced and colorcoded, singletons are shown in black.

ROTOWIRE	#Supp	#Contra	Gram	Coher	Concis
Gold	2.98*	0.28*	4.07*	3.33	-10.74*
TEMPL	6.98*	0.21*	-3.70*	-3.33*	17.78*
NCP+CC	4.90	0.90	-3.33*	-3.70*	-3.70
ENT	4.77	0.80	2.96	3.70	-3.33

MLB	#Supp	#Contra	Gram	Coher	Concis
Gold	2.81	0.15*	1.24*	3.48*	-9.33*
TEMPL	3.98*	0.04*	-10.67*	-7.30*	8.43*
ED+CC	3.24*	0.40	0.22*	-0.90*	-2.47*
NCP+CC	2.86	0.88*	0.90*	-1.35*	-1.80*
ENT	2.86	0.52	8.31	6.07	5.39

Table 5: Average number of supporting and contradicting facts in game summaries and *best-worst scaling* evaluation (higher is better) on ROTOWIRE and MLB datasets. Systems significantly different from ENT are marked with an asterisk * (using a one-way ANOVA with posthoc Tukey HSD tests; $p \leq 0.05$).

Thanks