Bi-directional attention flow for machine comprehension

ICLR 2017

Minjoon Seo1∗ Aniruddha Kembhavi2 Ali Farhadi1,2 Hananneh Hajishirzi1

Soojung Kim

what is MC/MRC (Machine Comprehension)?

Answering a query about a given context paragraph

Input:

How many touchdowns did T.J. Houshmandzadeh have? question

context paragraph

Still searching for their first win, the Bengals flew to Texas Stadium for a Week 5 interconference duel with the Dallas Cowboys. In the first quarter, Cincinnati trailed early as Cowboys kicker Nick Folk got a 30-yard field goal, along with RB Felix Jones getting a 33-yard TD run. In the second quarter, Dallas increased its lead as QB Tony Romo completed a 4-yard TD pass to TE Jason Witten. The Bengals would end the half with kicker Shayne Graham getting a 41-yard and a 31-yard field goal. In the third quarter, Cincinnati tried to rally as QB Carson Palmer completed an 18-yard TD pass to WR T. J. Houshmandzadeh. In the fourth quarter, the Bengals got closer as Graham got a 40-yard field goal, yet the Cowboys answered with Romo completing a 57-yard TD pass to WR Terrell Owens. Cincinnati tried to come back as Palmer completed a 10-yard TD pass to Houshmandzadeh (with a failed 2-point conversion), but Dallas pulled away with Romo completing a 15-yard TD pass to WR Patrick Crayton.

Output:

answer

Prediction [small, 60 million parameters]: two

Prediction [large, 770 million parameters]: 2

Correct Answer [only for our examples]: 2

before start

동일한 task에 대해서 다른 방법으로 접근

arxiv.org › cs ▾ 이 페이지 번역하기

Bidirectional Attention Flow for Machine Comprehension

In this paper we introduce the **Bi-Directional Attention Flow** (BIDAF) network, a multi-stage hierarchical process that represents the context at different levels of ...

M Seo 저술 - 2016 - 1017회 인용 - 관련 학술자료

Abstract

a multi-stage hierarchical process that represents the context at different levels of granularity and uses bi-directional attention flow mechanism to obtain a query-aware context representation without early summarization.

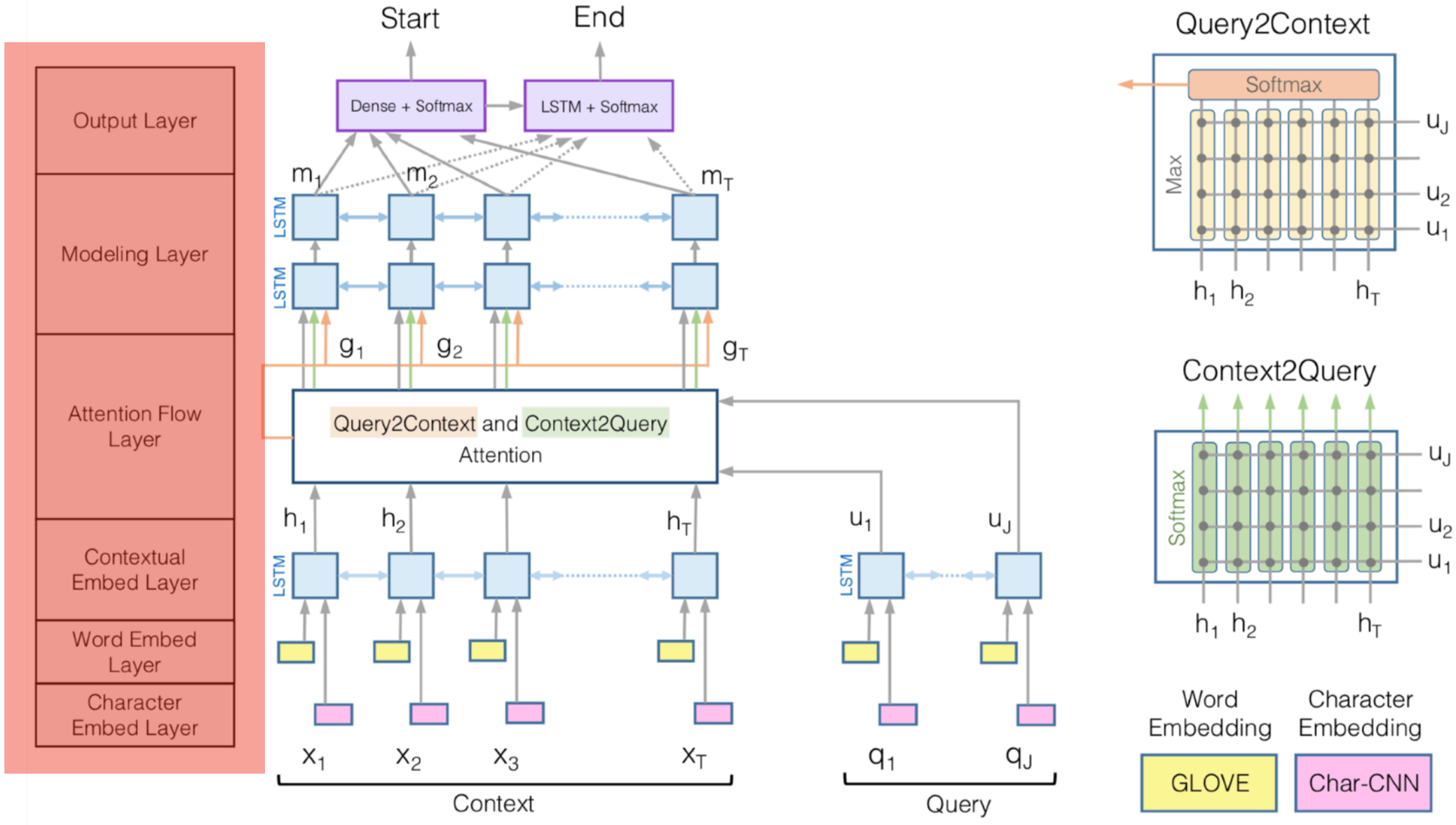


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Encoder module

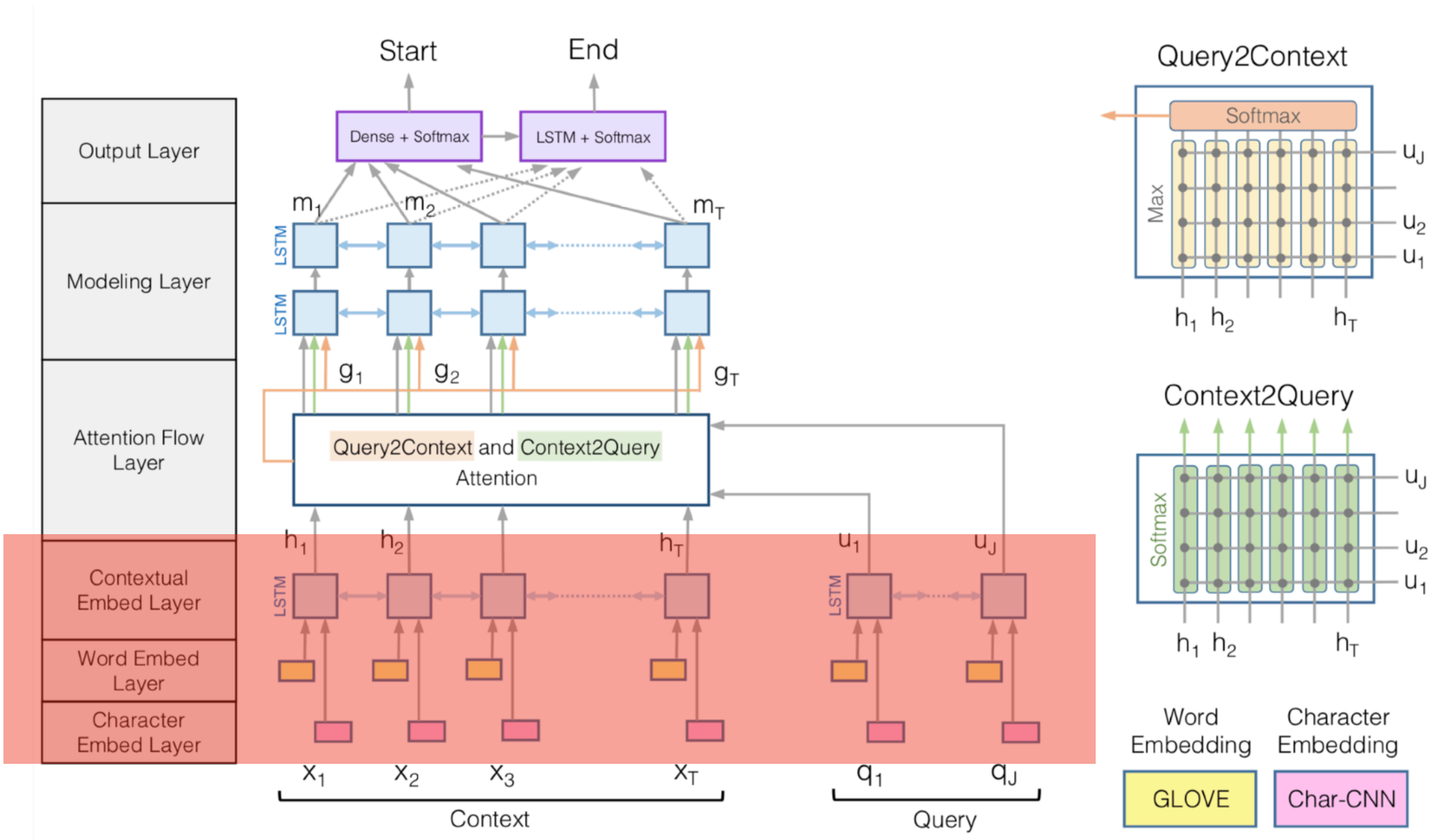


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Attention Flow

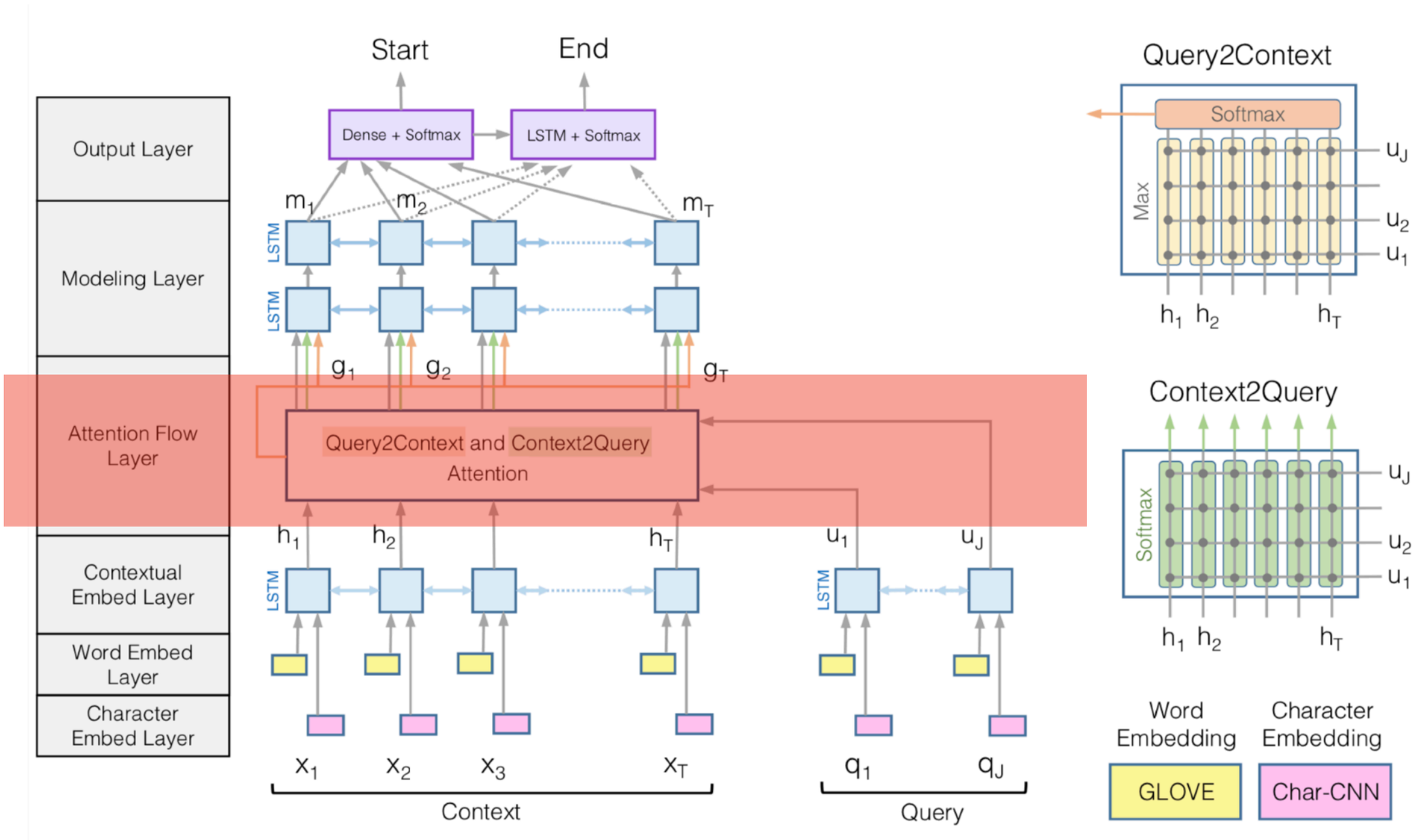


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Answer prediction

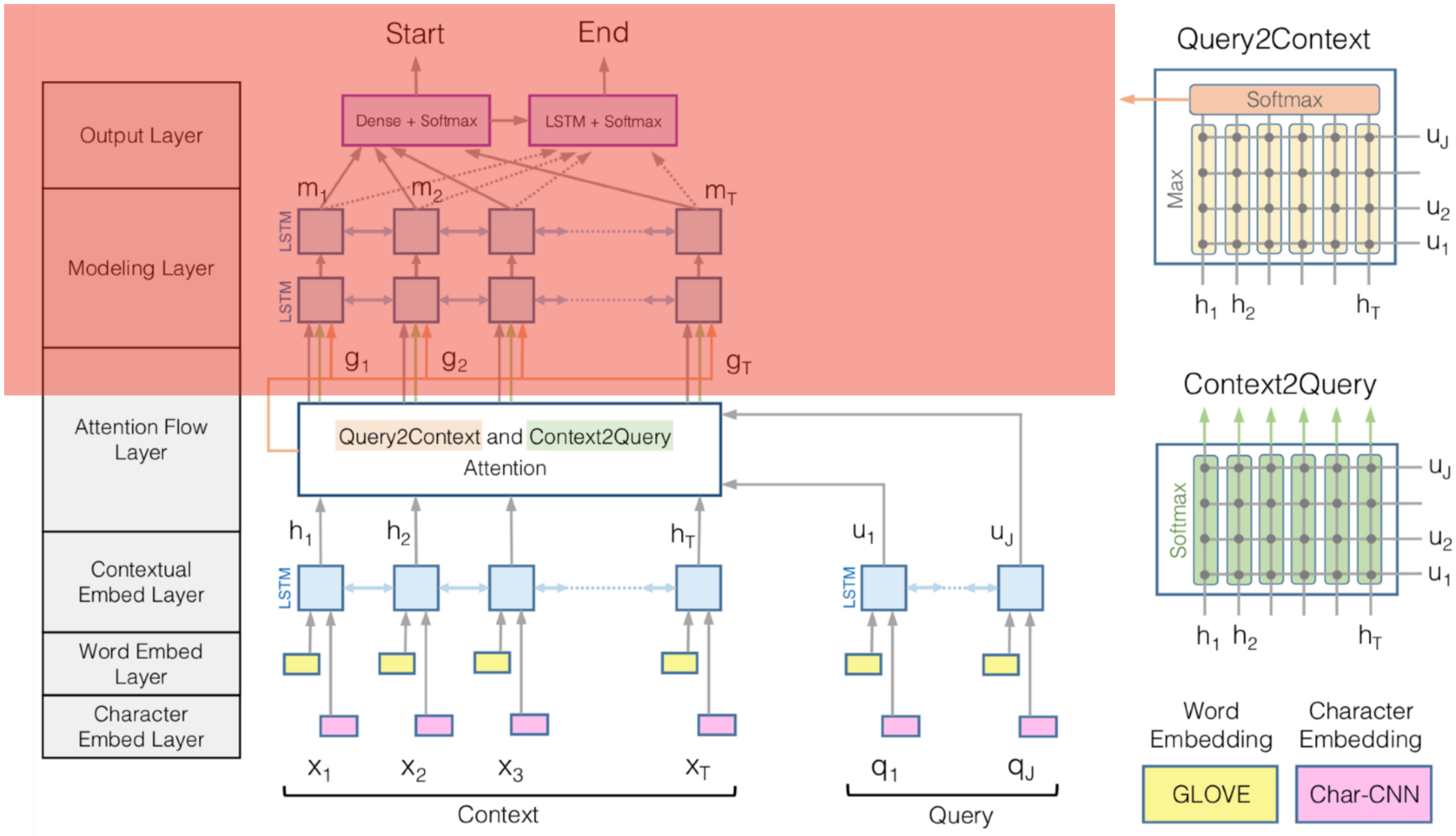


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Introduction

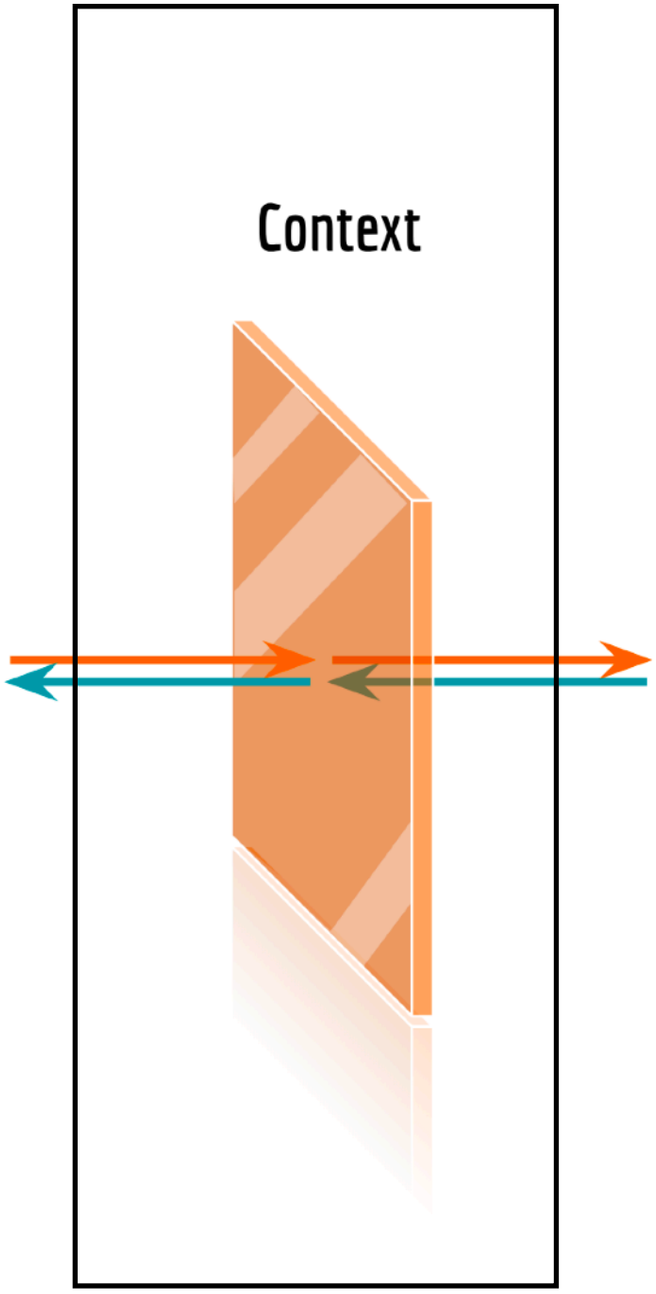
In previous works,
- Summarizing the context into fixed-size vector
- Temporally dynamic
- Uni-directional

question answering (QA) →
question generation (QG) →

Question

What is the function of lines of longitude and latitude?

Fixed-size



Answer

allow us to find an absolute location of a point on earth.

Introduction

In BiDAF, - **Not used to summarize the context paragraph into a fixed-size vector**
- Memory-less attention mechanism
- Bi-directional

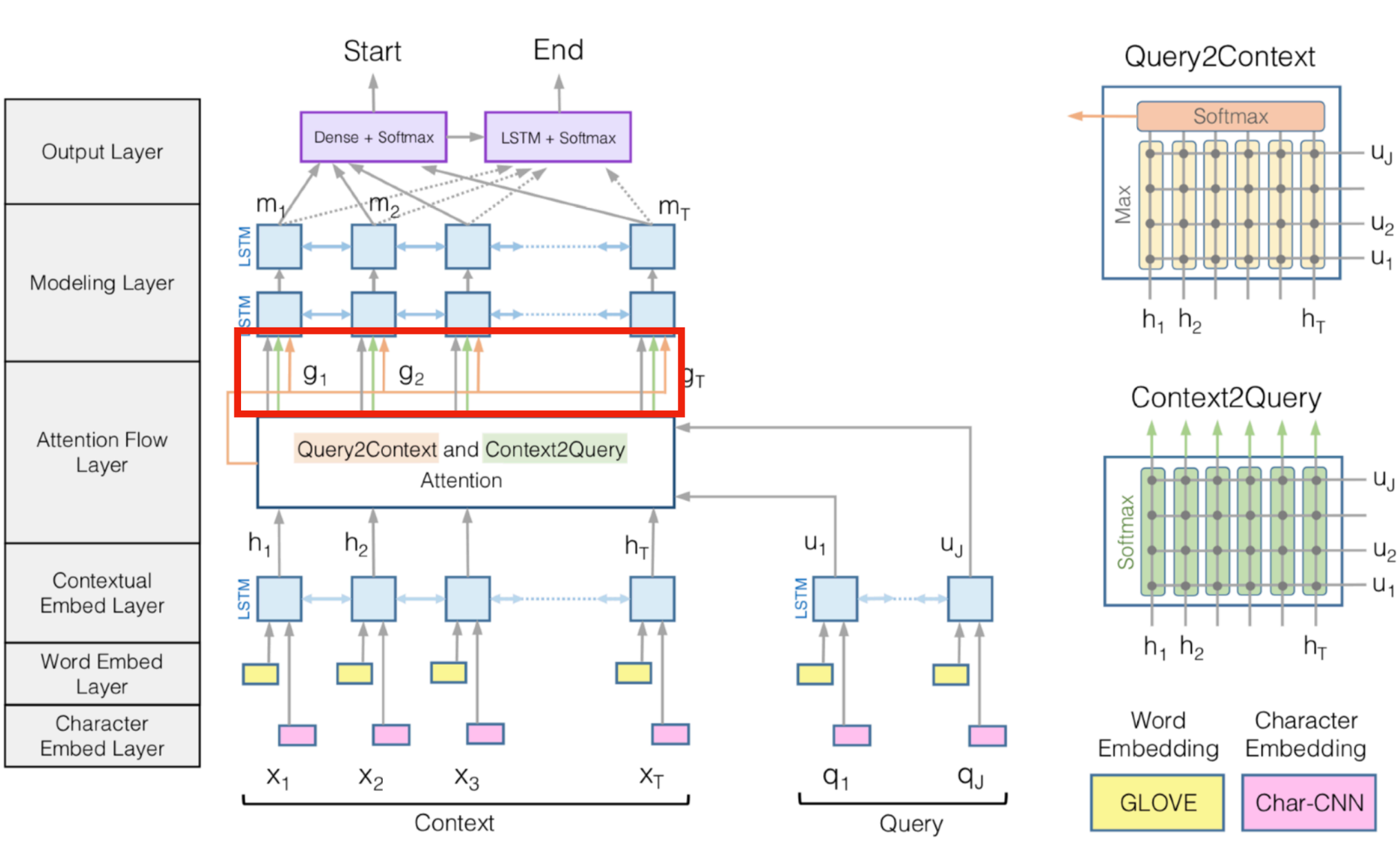


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Introduction

In BiDAF, - Not used to summarize the context paragraph into a fixed-size vector
- **Memory-less attention mechanism**
- Bi-directional

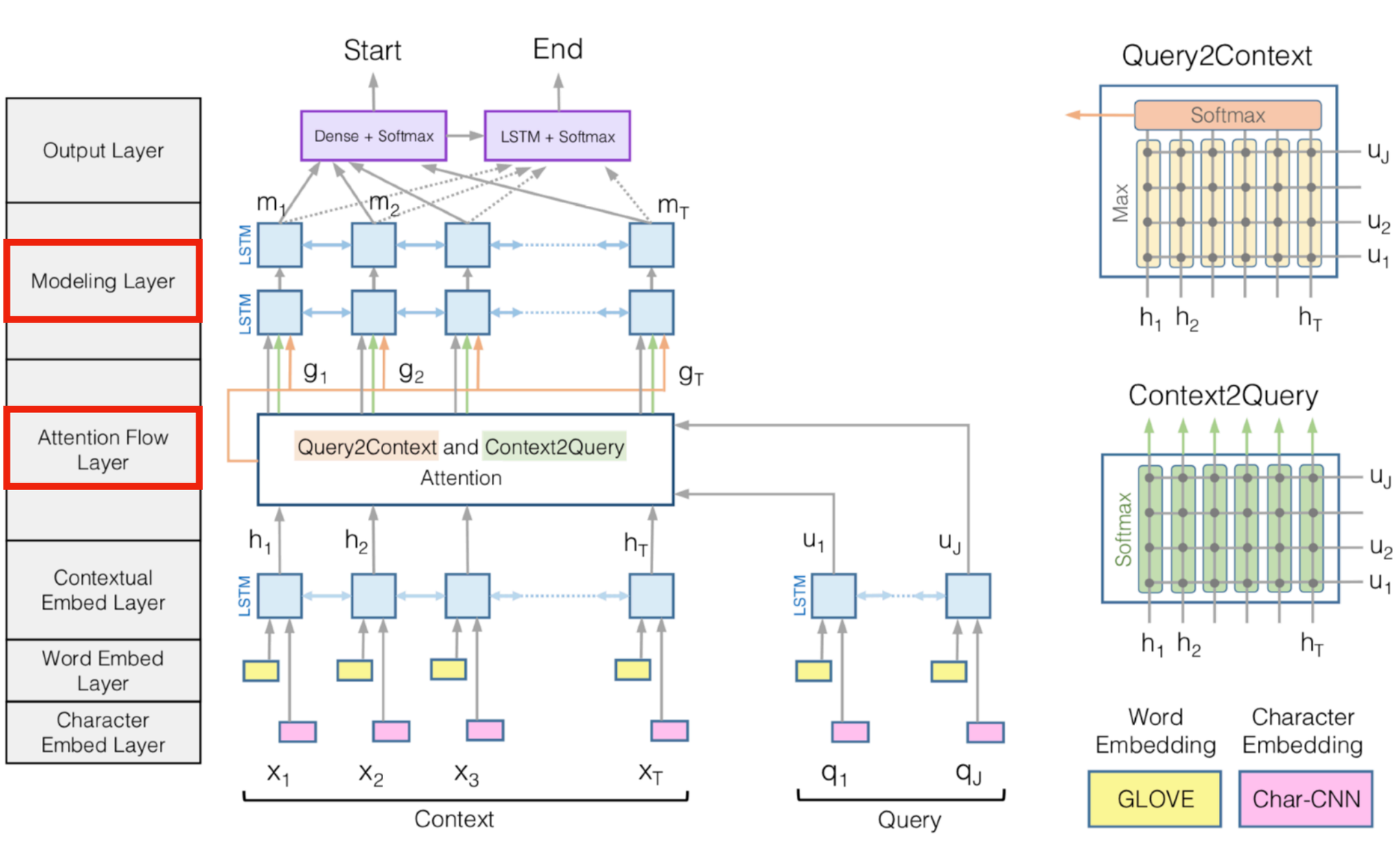


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Introduction

In BiDAF, - Not used to summarize the context paragraph into a fixed-size vector
- Memory-less attention mechanism
- **Bi-directional**

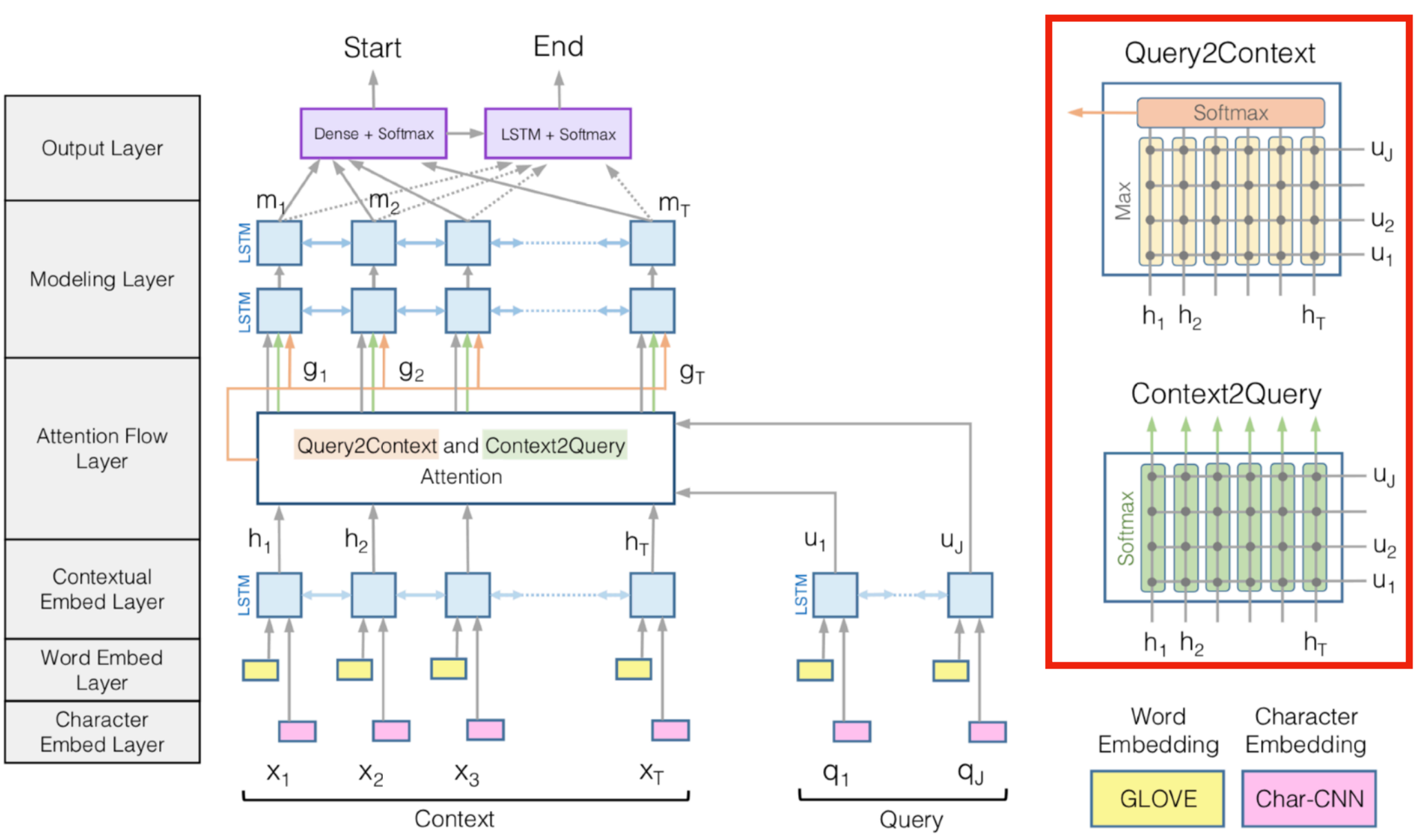


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Model

Encoder module

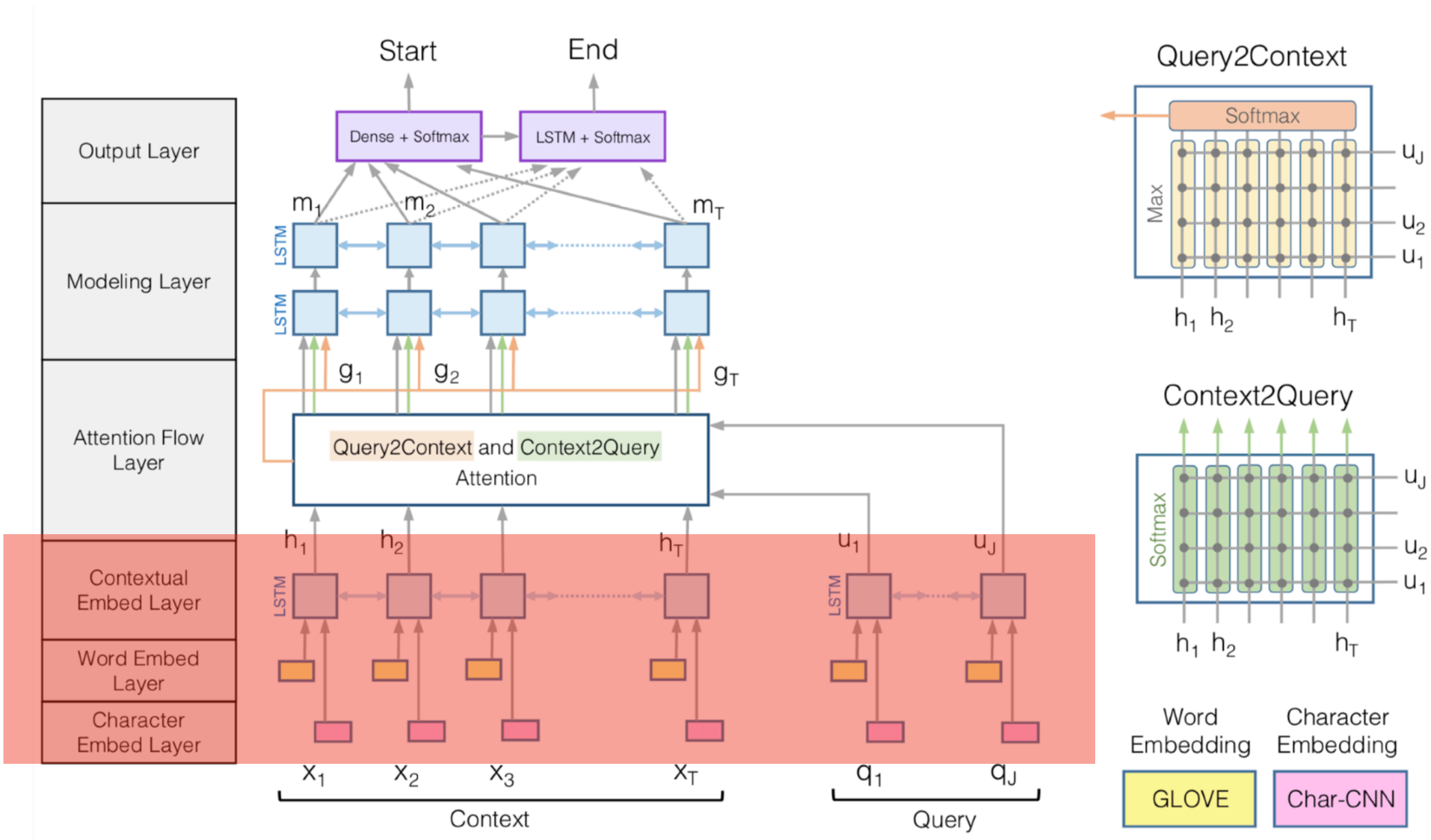
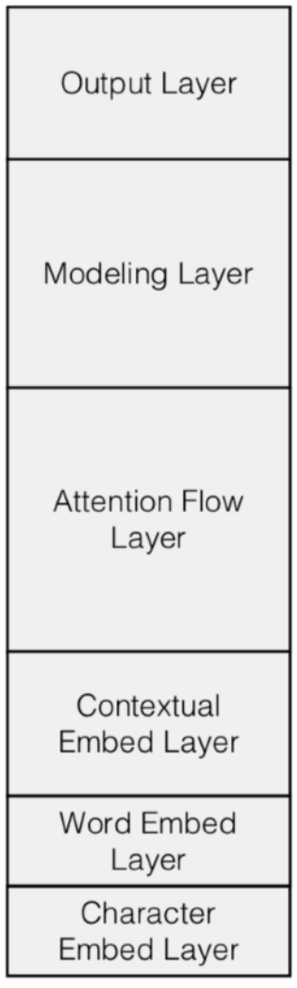


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Encoder module (1/3)

Embedding Layer:
Character embedding layer maps each word to a high-dimensional vector space using character-level CNNs.



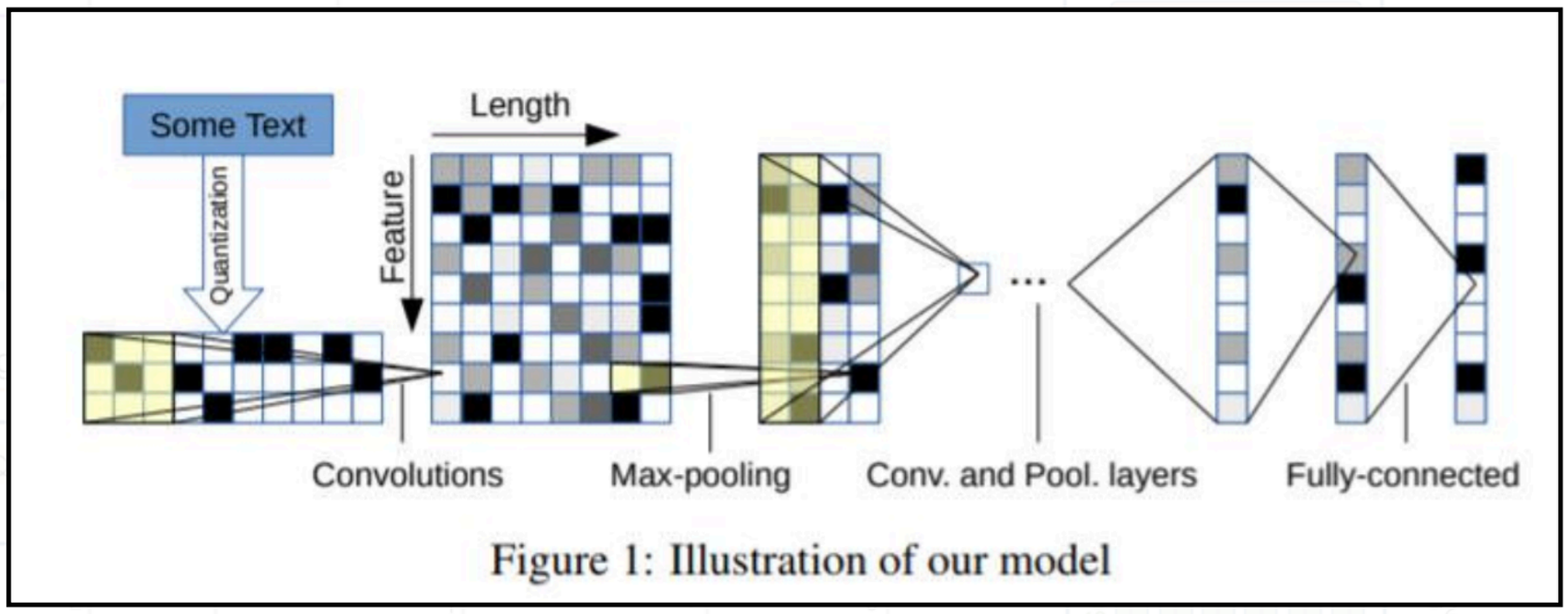


Figure 1: Illustration of our model

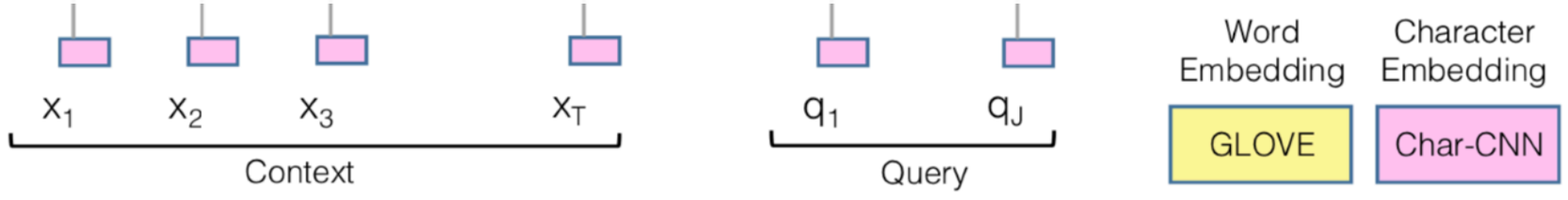


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Encoder module (2/3)

Embedding Layer:
Word embedding layer maps each word to a vector space using a pre-trained word embedding model. e.g. GLOVE

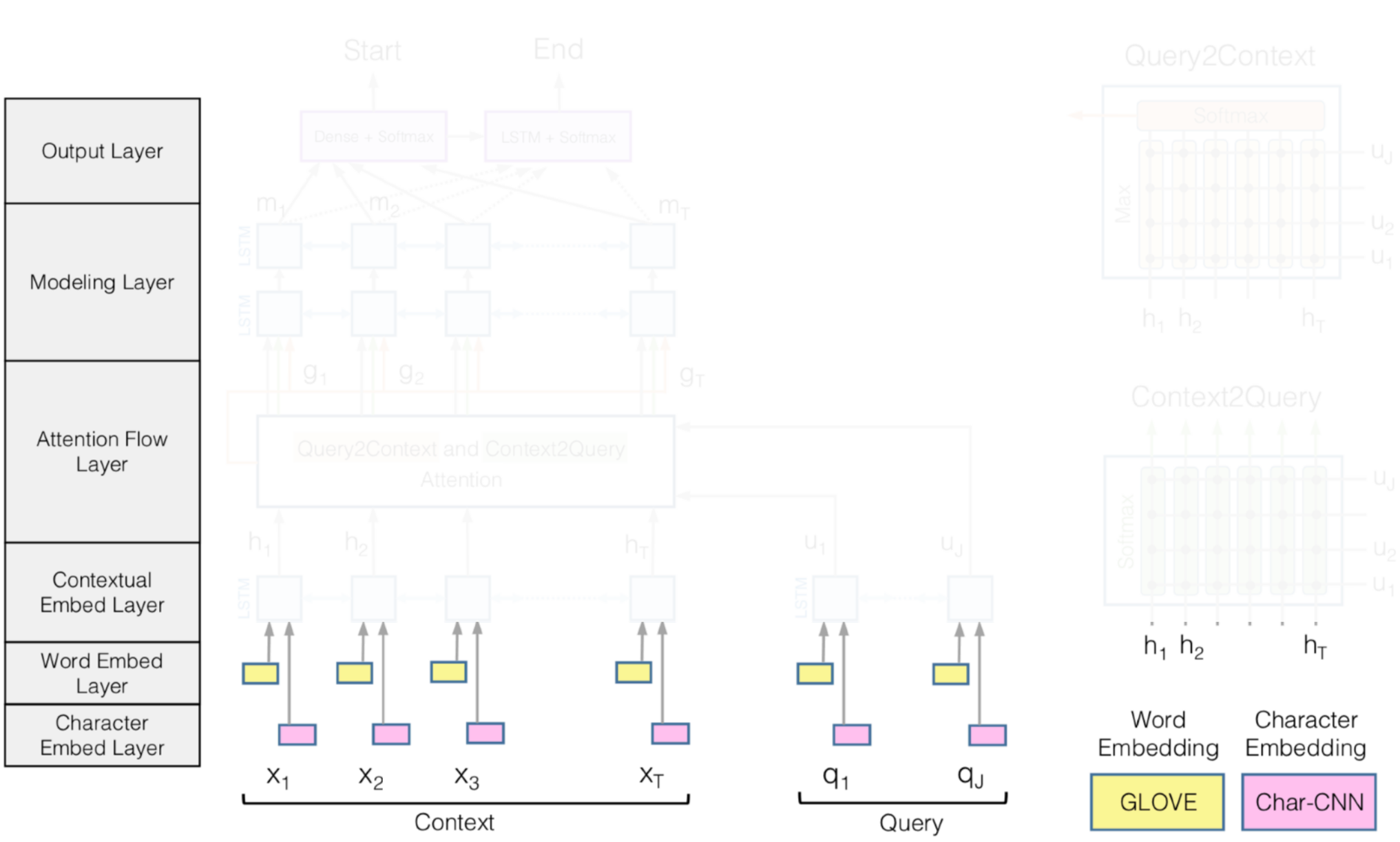


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Encoder module (3/3)

Embedding Layer:
Contextual embedding layer utilizes contextual cues from surrounding words to refine the embedding of the words. These first three layers are applied to both the query and context.

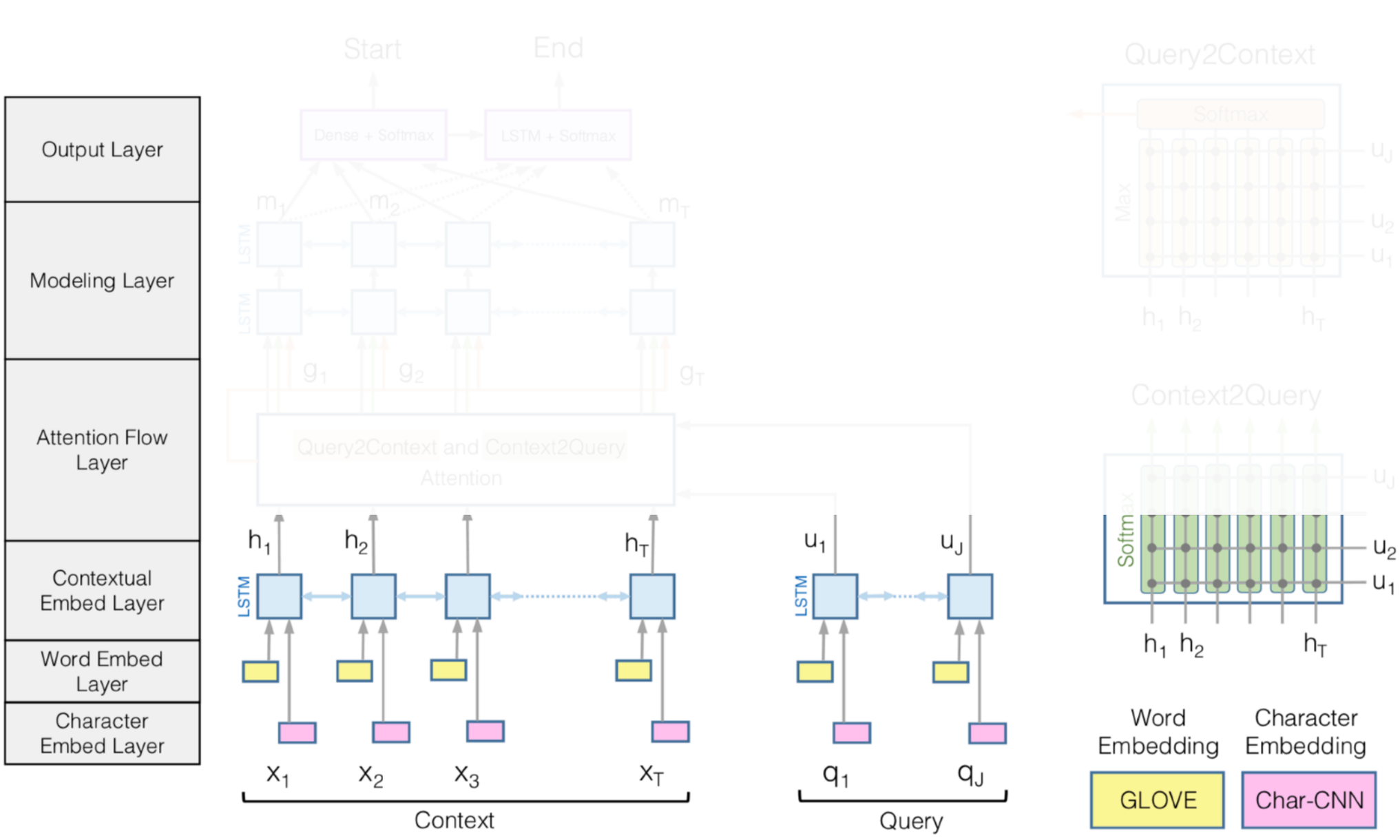
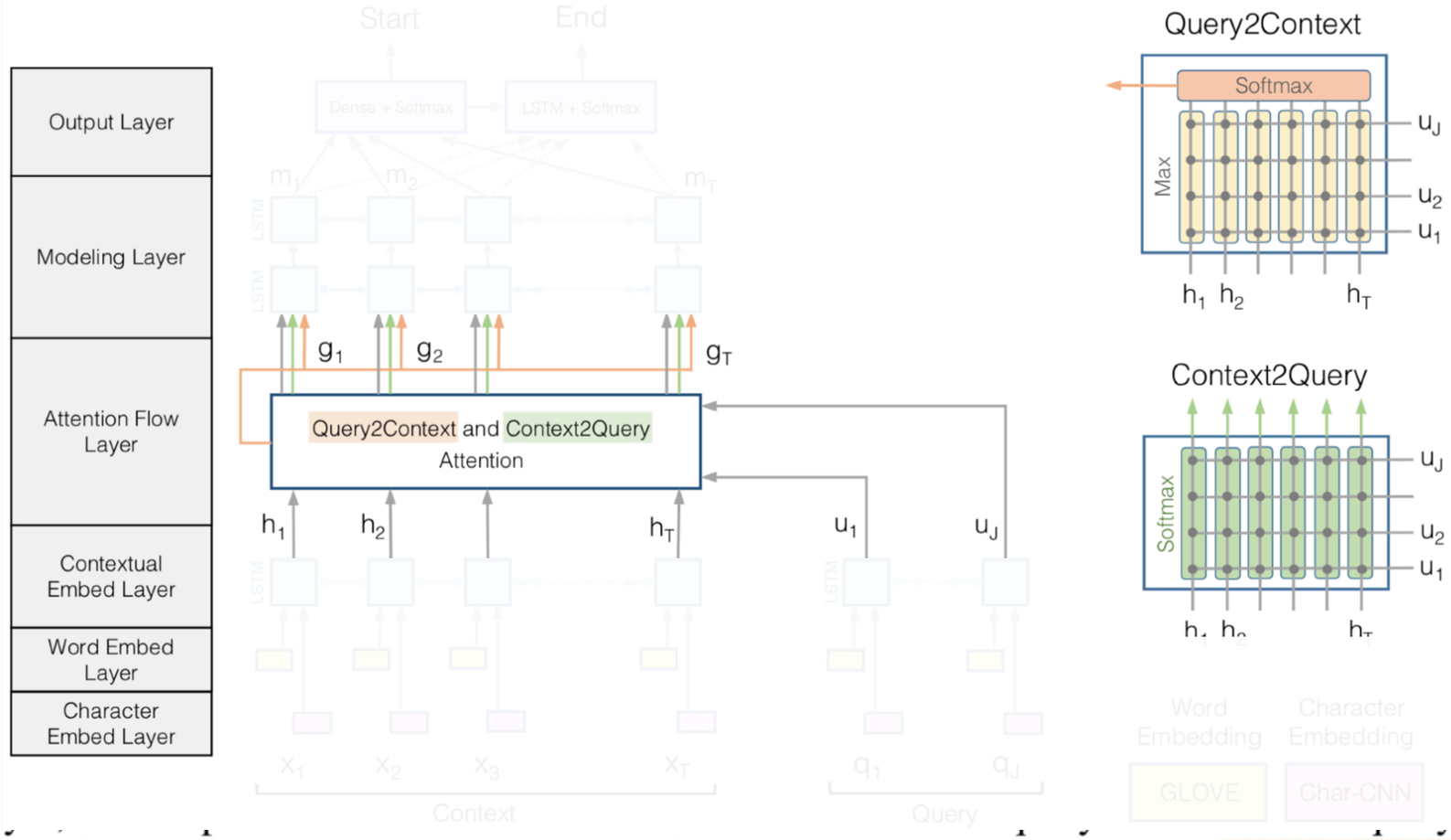


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Attention Flow

Attention Flow Layer:
Linking and fusing information from the context and the query words.



context. Both of these attentions, which will be discussed below, are derived from a shared similarity matrix, $\mathbf{S} \in \mathbb{R}^{T \times J}$, between the contextual embeddings of the context ($\mathbf{H}$) and the query ($\mathbf{U}$), where $\mathbf{S}_{tj}$ indicates the similarity between t-th context word and j-th query word. The similarity matrix is computed by

$$\mathbf{S}_{tj} = \alpha(\mathbf{H}_{:t}, \mathbf{U}_{:j}) \in \mathbb{R} \tag{1}$$

Attention Flow (1/2)

Context2Query attention signifies which query words are most relevant to each context word.

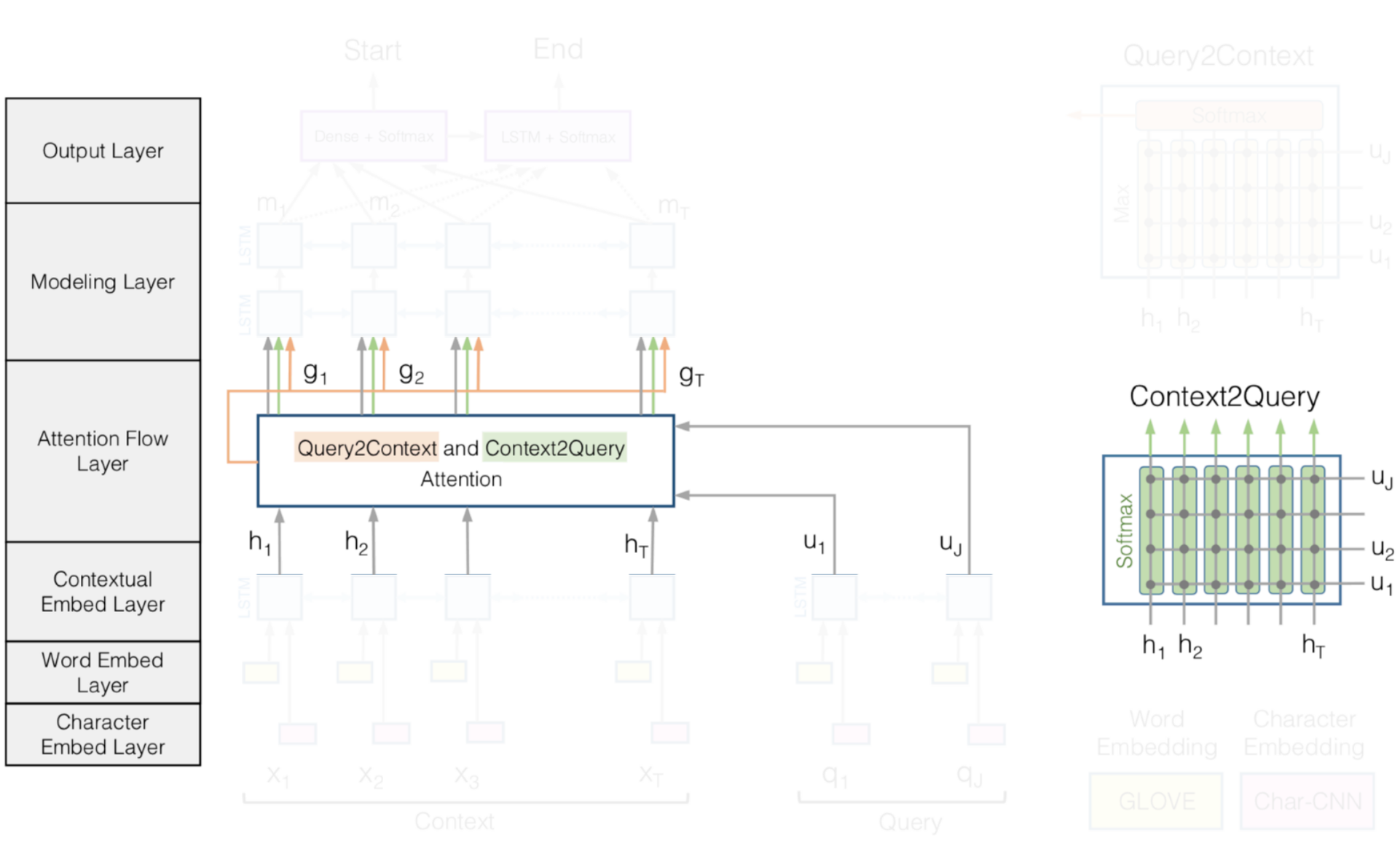


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Attention Flow (2/2)

Query2Context attention signifies which context words have the closest similarity to one of the query words and are hence critical for answering the query.

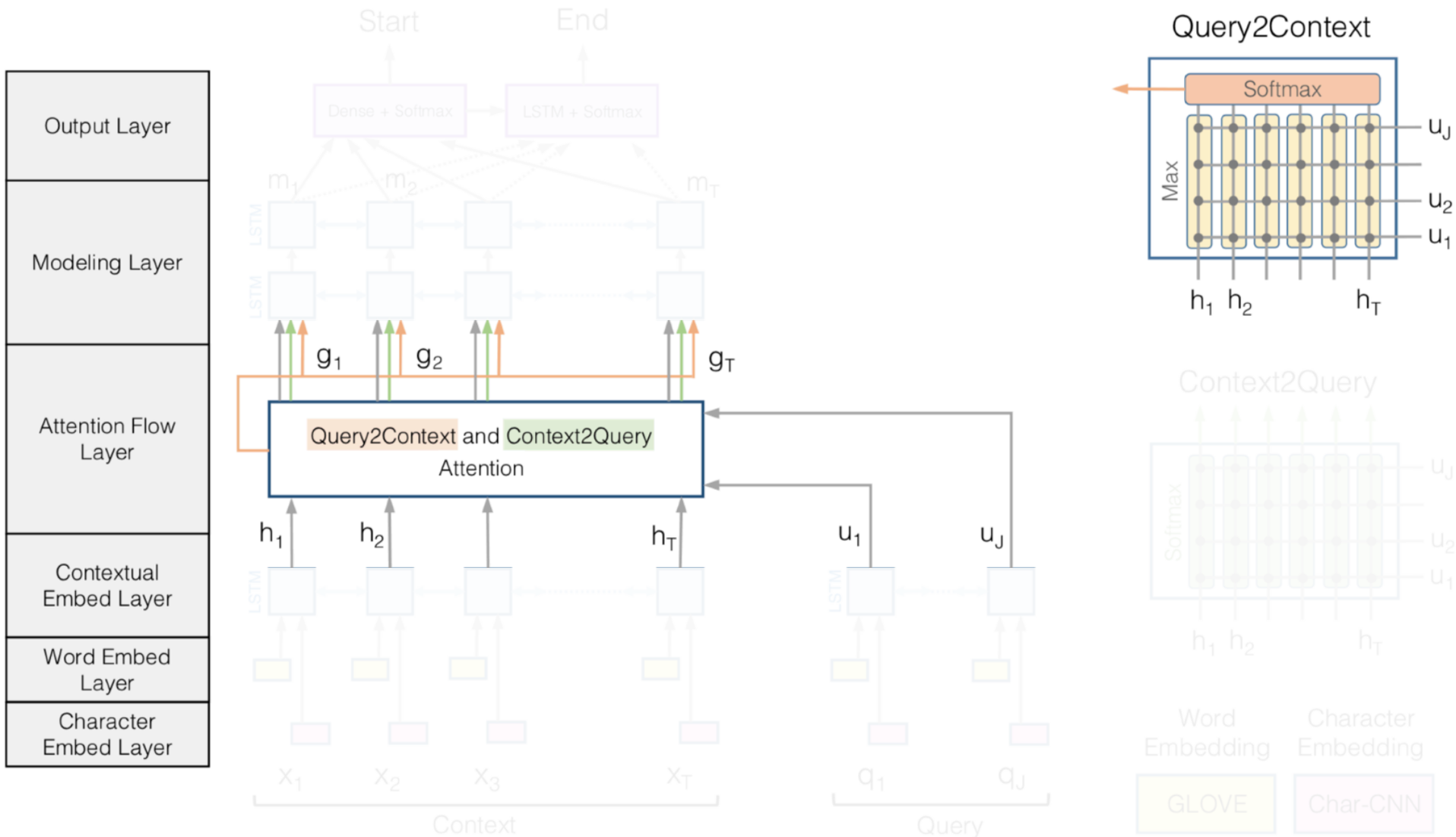


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Model

Modeling Layer:
Recurrent Neural Network to scan the context.

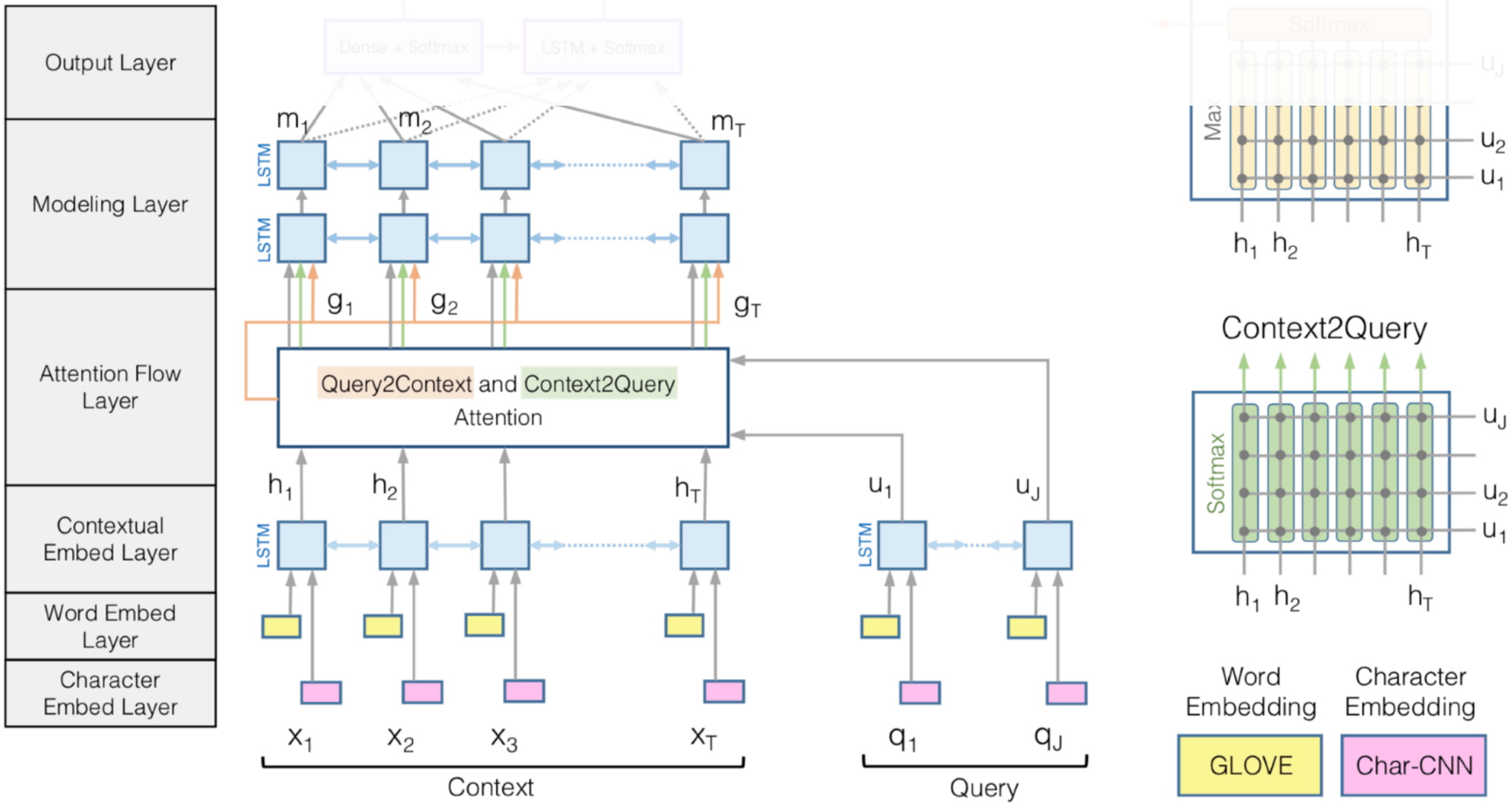


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Model

Output Layer:
provides an answer to the query.

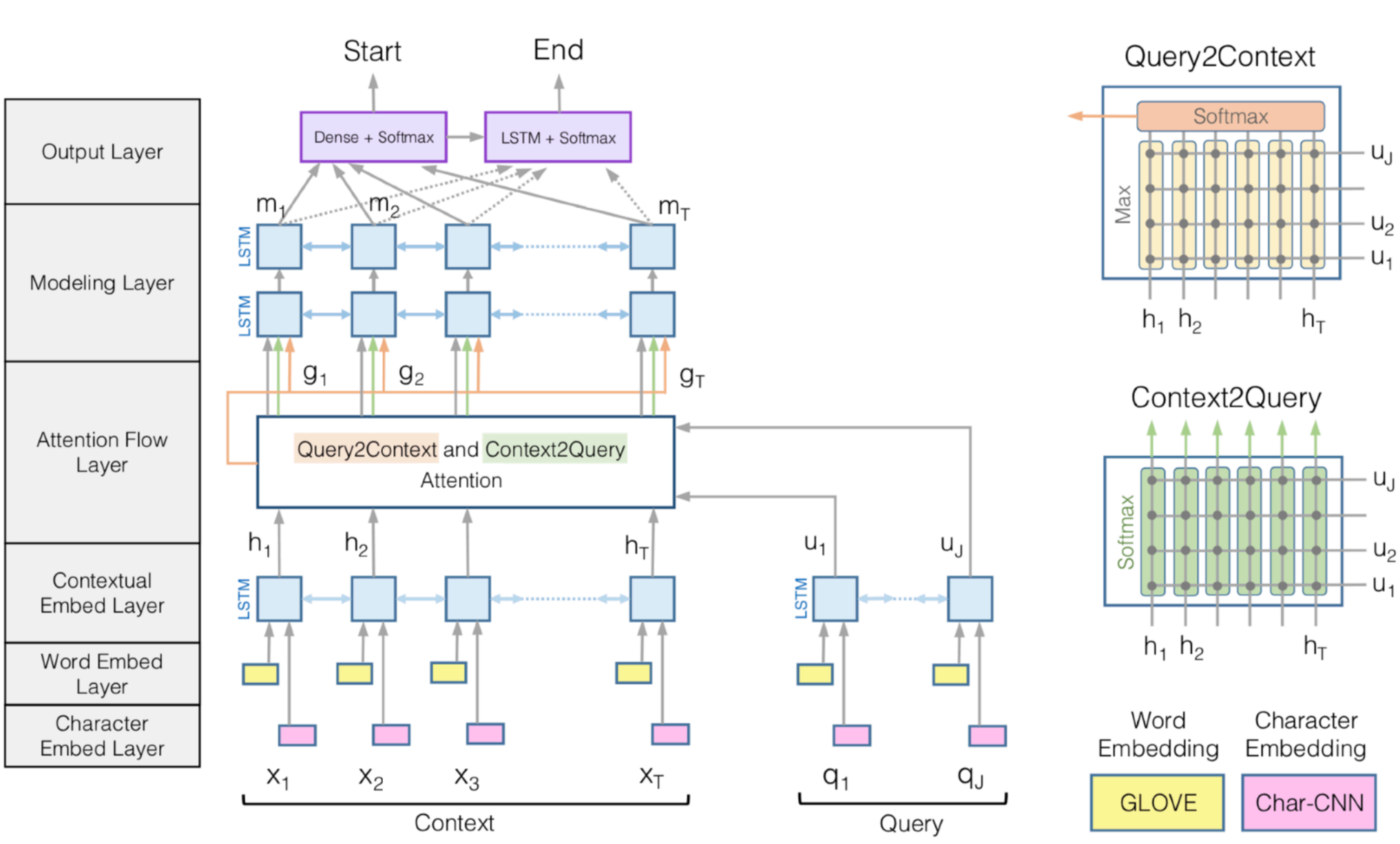


Figure 1: BiDirectional Attention Flow Model *(best viewed in color)*

Result

	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BIDAF (Ours)	68.0	77.3	**73.3**	**81.1**

(a) Results on the SQuAD test set

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

(b) Ablations on the SQuAD dev set

Visualization (1/2)

Layer	Query	Closest words in the Context using cosine similarity
Word	When	when, When, After, after, He, he, But, but, before, Before
Contextual	When	When, when, 1945, 1991, 1971, 1967, 1990, 1972, 1965, 1953
Word	Where	Where, where, It, IT, it, they, They, that, That, city
Contextual	Where	where, Where, Rotterdam, area, Nearby, location, outside, Area, across, locations
Word	Who	Who, who, He, he, had, have, she, She, They, they
Contextual	Who	who, whose, whom, Guiscard, person, John, Thomas, families, Elway, Louis
Word	city	City, city, town, Town, Capital, capital, district, cities, province, Downtown
Contextual	city	city, City, Angeles, Paris, Prague, Chicago, Port, Pittsburgh, London, Manhattan
Word	January	July, December, June, October, January, September, February, April, November, March
Contextual	January	January, March, December, August, December, July, July, July, March, December
Word	Seahawks	Seahawks, Broncos, 49ers, Ravens, Chargers, Steelers, quarterback, Vikings, Colts, NFL
Contextual	Seahawks	Seahawks, Broncos, Panthers, Vikings, Packers, Ravens, Patriots, Falcons, Steelers, Chargers
Word	date	date, dates, until, Until, June, July, Year, year, December, deadline
Contextual	date	date, dates, December, July, January, October, June, November, March, February

Table 2: Closest context words to a given query word, using a cosine similarity metric computed in the Word Embedding feature space and the Phrase Embedding feature space.

Visualization (2/2)

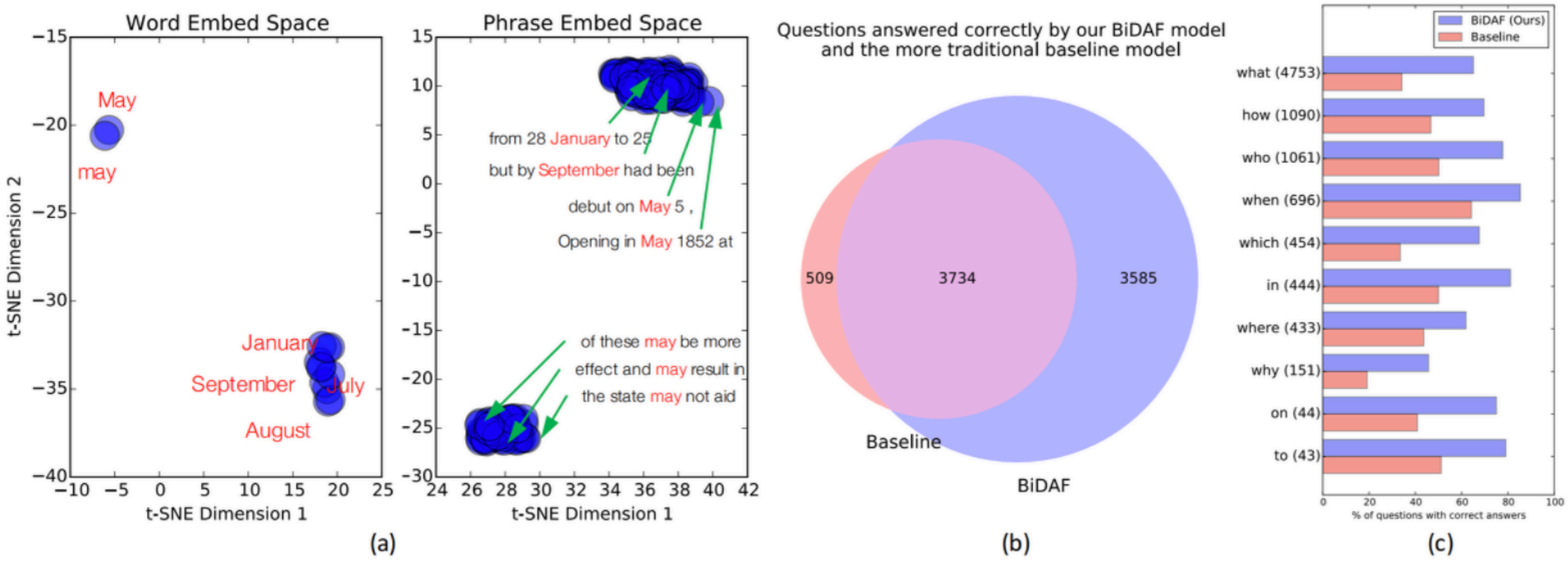


Figure 2: (a) t-SNE visualizations of the *months* names embedded in the two feature spaces. The contextual embedding layer is able to distinguish the two usages of the word *May* using context from the surrounding text. (b) Venn diagram of the questions answered correctly by our model and the *more traditional* baseline (Rajpurkar et al., 2016). (c) Correctly answered questions broken down by the 10 most frequent first words in the question.

Thanks