Utilizing lexicon in our NLU

Why do we want to utilize lexicon?

- Let's take ATIS dataset
- It has finite set of cities in training
- Will the model work for a new city?
- We have **a list of all cities**, why not use it?

- Another example
- Imagine you need to fill a slot "music artist"
- We have **all music artists** in the database like musicbrainz.org
- How can we use it?

Let's add lexicon features to input words

- Let's **match every n-gram** of input text against entries in our lexicon

Take me to San Francisco

- A match is successful when **the n-gram matches the prefix or postfix** of an entry and is at least half the length of the entry

Matches:

"San" → "San Antonio"
"San" → "San Francisco"
"San Francisco" → "San Francisco"

- When there are multiple **overlapping matches**:
  - Prefer **exact** matches over partial
  - Prefer **longer** matches over shorter
  - Prefer **earlier** matches in the sentence over later

https://arxiv.org/pdf/1511.08308v4.pdf

Matches encoding

We will use **BIOES** coding (Begin, Inside, Outside, End, Single)

- B – if token matches the beginning of some entity
- B, I – if two tokens match as prefix
- I, E – if two tokens match as postfix
- S – if matched single token entity
- …

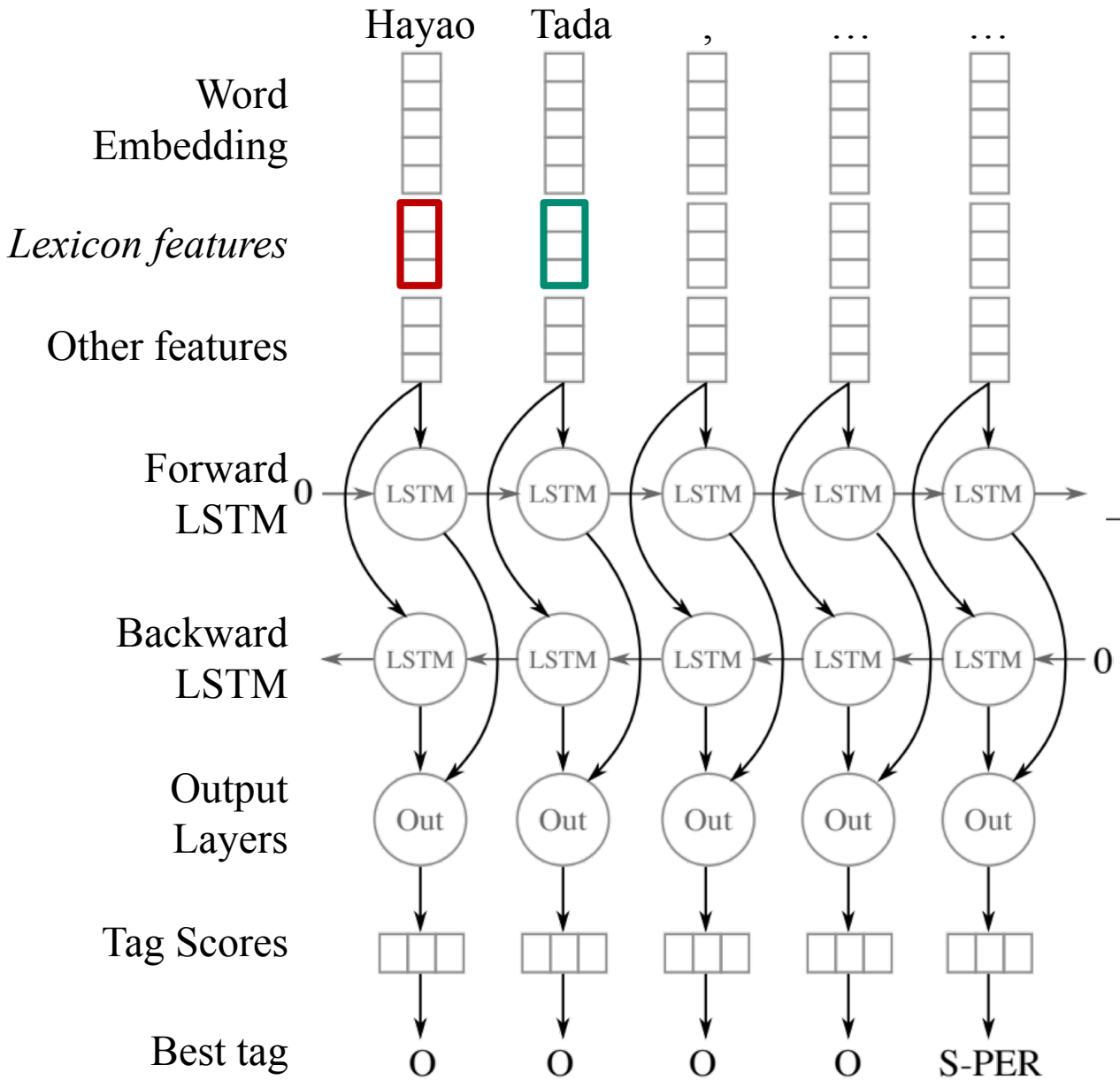
Example for 4 lexicon dictionaries:

Text	Hayao	Tada	,	commander	of	the	Japanese	North	China	Area	Army
LOC	–	–	–	–	–	B	I	–	S	–	–
MISC	–	–	–	S	B	B	I	S	S	S	S
ORG	–	–	–	–	–	B	I	B	I	I	E
PERS	B	E	–	–	–	–	–	–	S	–	–

B, I, O, E, S are later encoded as one-hot vectors

https://arxiv.org/pdf/1511.08308v4.pdf

Adding these features to our model



These're our lexicon features

Text	Hayao	Tada	,
LOC	–	–	–
MISC	–	–	–
ORG	–	–	–
PERS	B	E	–

https://arxiv.org/pdf/1511.08308v4.pdf

Does lexicon help?

CoNLL-2003 Named Entity Recognition task:

Model	CoNLL-2003		
	Prec.	Recall	F1
FFNN + emb + caps + lex	89.54	89.80	89.67 (± 0.24)
BLSTM	80.14	72.81	76.29 (± 0.29)
BLSTM-CNN	83.48	83.28	83.38 (± 0.20)
BLSTM-CNN + emb	90.75	91.08	90.91 (± 0.20)
BLSTM-CNN + emb + lex	91.39	**91.85**	**91.62** (± 0.33)

Yes, it does!

Training details

- You can **sample** your **lexicon** dictionaries so that your model learns the context of entities as well as lexicon features
- This procedure helps **to detect unknown entities at testing**

- You can **augment** your dataset replacing slot values with values from the same lexicon:

Take me to **San Francisco**

Take me to **Washington**

Summary

- You can add lexicon features to further improve your NLU
- In the next video we'll take a look at Dialog Manager (DM)