Character-based Deep Learning Models for Token and Sentence Segmentation

Alymzhan Toleu[a,b], Gulmira Tolegen[a,b],
Aibek Makazhanov[a]

(a) National Laboratory Astana, Nazarbayev University, Astana
(b) Tsinghua University, Beijing

Outline

Definitions

Tokenization

- `Input:` I can't do it. Not again!
- `Output:` [I][can]['t][do][it][.][Not][again][!]

Definitions

Tokenization

- `Input:` I can't do it. Not again!
- `Output:` [I][can]['t][do][it][.][Not][again][!]

Sentence segmentation

- `Input:` I can't do it. Not again!
- `Output:` [I can't do it.][Not again !]

Definitions

Tokenization

- `Input:` I can't do it. Not again!
- `Output:` [I][can]['t][do][it][.][Not][again][!]

Sentence segmentation

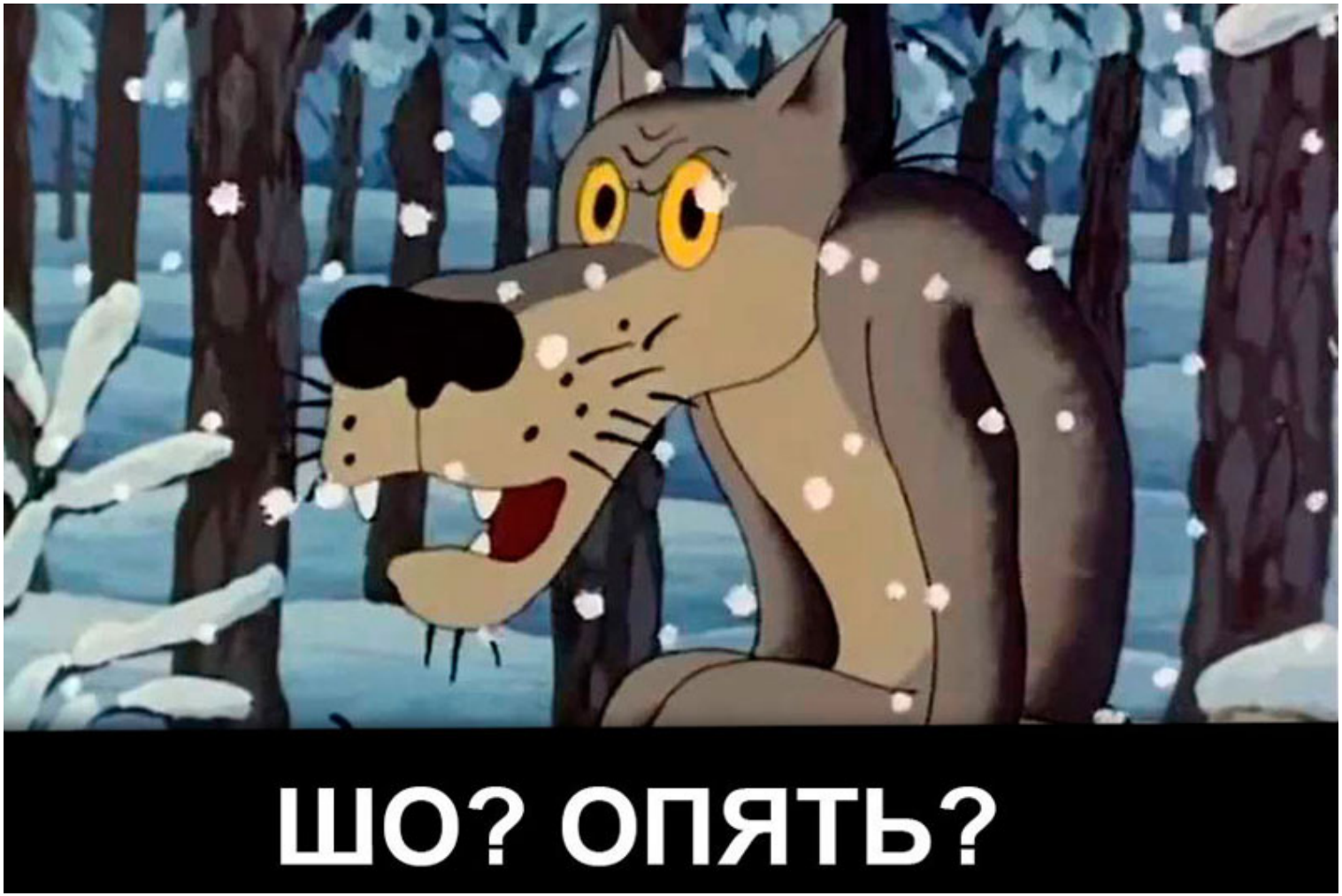
- `Input:` I can't do it. Not again!
- `Output:` [I can't do it.][Not again !]

TSS

- `Input:` I can't do it. Not again!
- `Output:` <[I][can]['t][do][it][.]><[Not][again][!]>

Motivation: Why bother? Isn't it solved?

Motivation: Why bother? Isn't it solved?



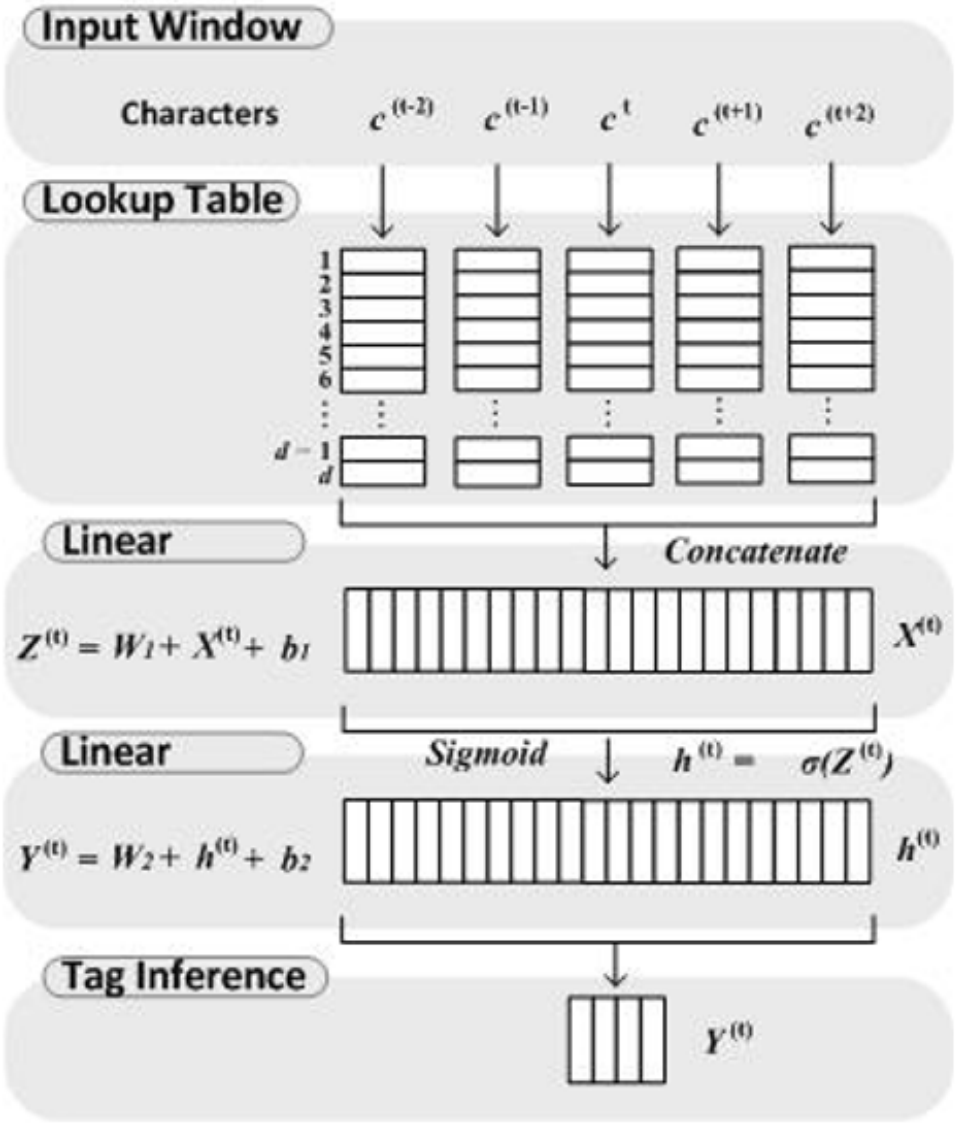
Solutions: Existing

- **Punkt** (Kiss and Strunk, 2006)
  unsupervised: lexicons, rules, regex, etc.;
- **Elephant** (Evang et al., 2013)
  supervised: hand-crafted features.

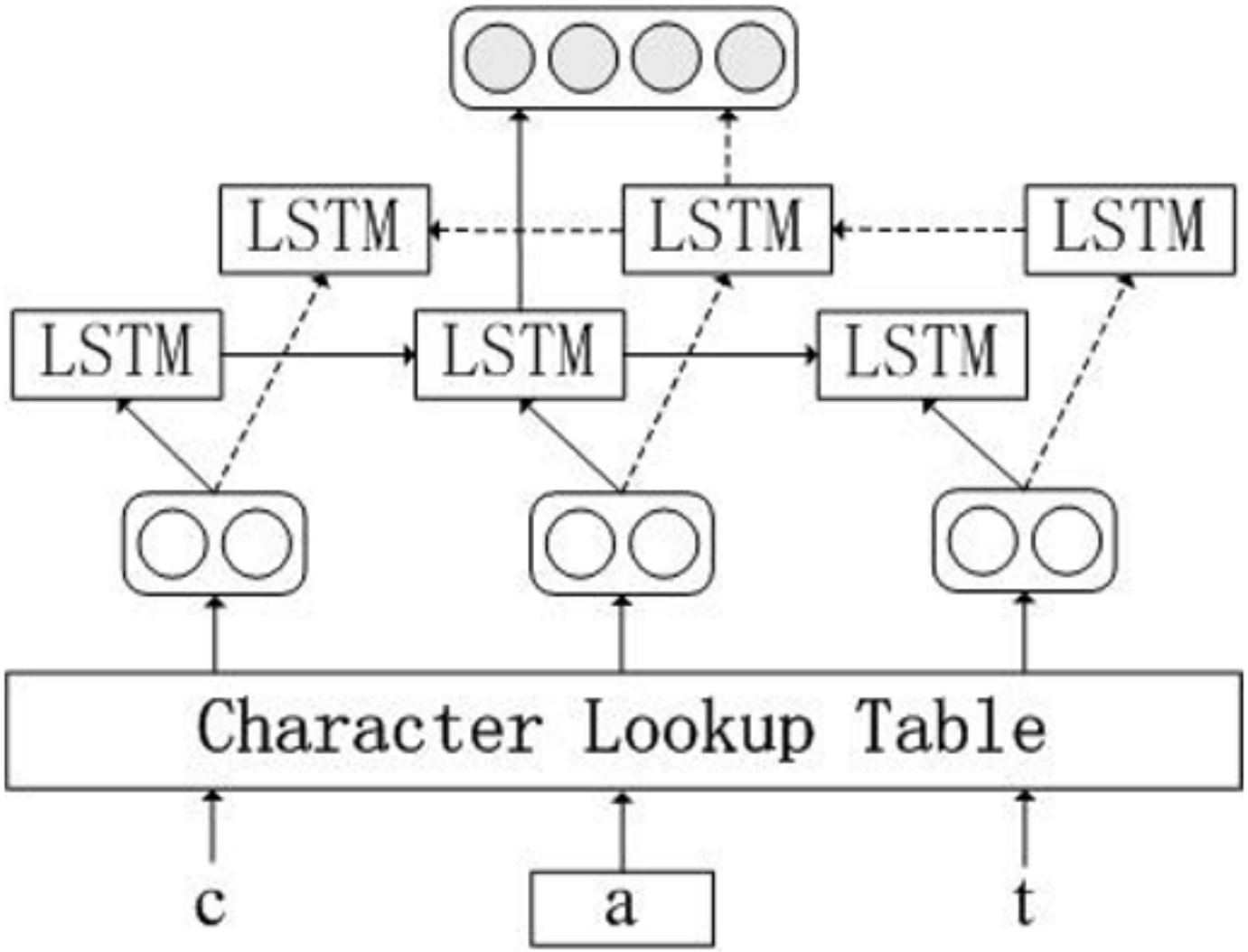
Solutions: IOB-labeling and tagging

```
ENG  input: I couldn’t do 100 sit-ups let alone 1 000.
     tags:  SOTIIIITIIOTIOTIIOTIIIIIIOTIIOTIIIIOTIIIIT
     tok-s: <I could n’t do 100 sit-ups let alone {1 000} .>

ITA  input: Grazie Italia!Ti ho dato l'oro.
     tags:  SIIIIIOTIIIIITSIOTIOTIIIOTITIIT
     tok-s: <Grazie Italia !><Ti ho dato l' oro .>

KAZ  input: Содан-ақ 2015ж. бұл көрсеткіш 4%-ға өскені белгілі.
     tags:  SIIIITIIOTIIITIOTIIOTIIIIIIIIOTTIIIOTIIIIIOTIIIIIIT
     tok-s: <Содан -ақ 2015 ж. бұл көрсеткіш 4 %-ға өскені белгілі .>
```

Solutions: Deep learning – General NN



Solutions: Deep learning – (bi)LSTM



Setup

- **Char embedding size**: 35;
- **Window size**: 9;
- **# hidden states**: 100;
- **Training time, epochs**: 300;
- **Evaluation**: P/R/F + error rate.

Data set

Table 1. Characteristics of the data sets.

Language	Domain	# sentences	# tokens
Kazakh	web/various	4 360	96,760
English	newswire	2 886	64,443
Italian	web/various	42 674	869,095

Results

Table 2. Evaluation results for English.

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**100**	**100**	**100**	99.92	99.82	99.87
LSTM	99.34	99.34	99.34	99.94	**99.86**	**99.90**
bi-LSTM	99.67	99.34	99.50	**99.95**	**99.86**	**99.90**

Results

Table 2. Evaluation results for English.

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**100**	**100**	**100**	99.92	99.82	99.87
LSTM	99.34	99.34	99.34	99.94	**99.86**	**99.90**
bi-LSTM	99.67	99.34	99.50	**99.95**	**99.86**	**99.90**

Table 3. Evaluation results for Italian.

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**99.28**	96.32	97.78	99.63	99.78	99.70
LSTM	99.00	96.27	97.62	99.52	99.71	99.61
bi-LSTM	99.25	**96.76**	**97.99**	**99.74**	**99.86**	**99.80**

Results

Table 2. Evaluation results for English.

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**100**	**100**	**100**	99.92	99.82	99.87
LSTM	99.34	99.34	99.34	99.94	**99.86**	**99.90**
bi-LSTM	99.67	99.34	99.50	**99.95**	**99.86**	**99.90**

Table 3. Evaluation results for Italian.

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**99.28**	96.32	97.78	99.63	99.78	99.70
LSTM	99.00	96.27	97.62	99.52	99.71	99.61
bi-LSTM	99.25	**96.76**	**97.99**	**99.74**	**99.86**	**99.80**

Table 4. Evaluation results for Kazakh.

	Sentence segmentation			Tokenization		
Models	Precision	Recall	F-measure	Precision	Recall	F-measure
NN	**92.70**	**99.44**	**95.95**	99.74	**99.44**	99.59
LSTM	92.43	97.95	95.11	99.58	99.43	99.50
bi-LSTM	92.20	99.25	95.60	**99.82**	99.40	**99.61**

Comparative evaluation

Table 5. Comparison with other systems.

	English		Italian	
Models	Sentence (F-measure)	Sent. + Tok. (error rate)	Sentence (F-measure)	Sent. + Tok. (error rate)
Punkt	98.51	-	98.34	-
Elephant	**100**	0.27	**99.51**	0.76
NN	**100**	0.05	97.78	0.12
LSTM	**100**	**0.03**	97.62	0.13
bi-LSTM	**100**	**0.03**	97.99	**0.07**

Conclusions and Future Work

- Models for joint TSS were build;
- No lexicons, rules, feature extraction – only data;
- In the future:
  - More data;
  - More training/tuning.