Dataset Mention Extraction and Classification

Animesh Prasad, Chenglei Si and Min-Yen Kan

National University of Singapore

{animesh, kanmy}@comp.nus.edu.sg

❖ Introduction

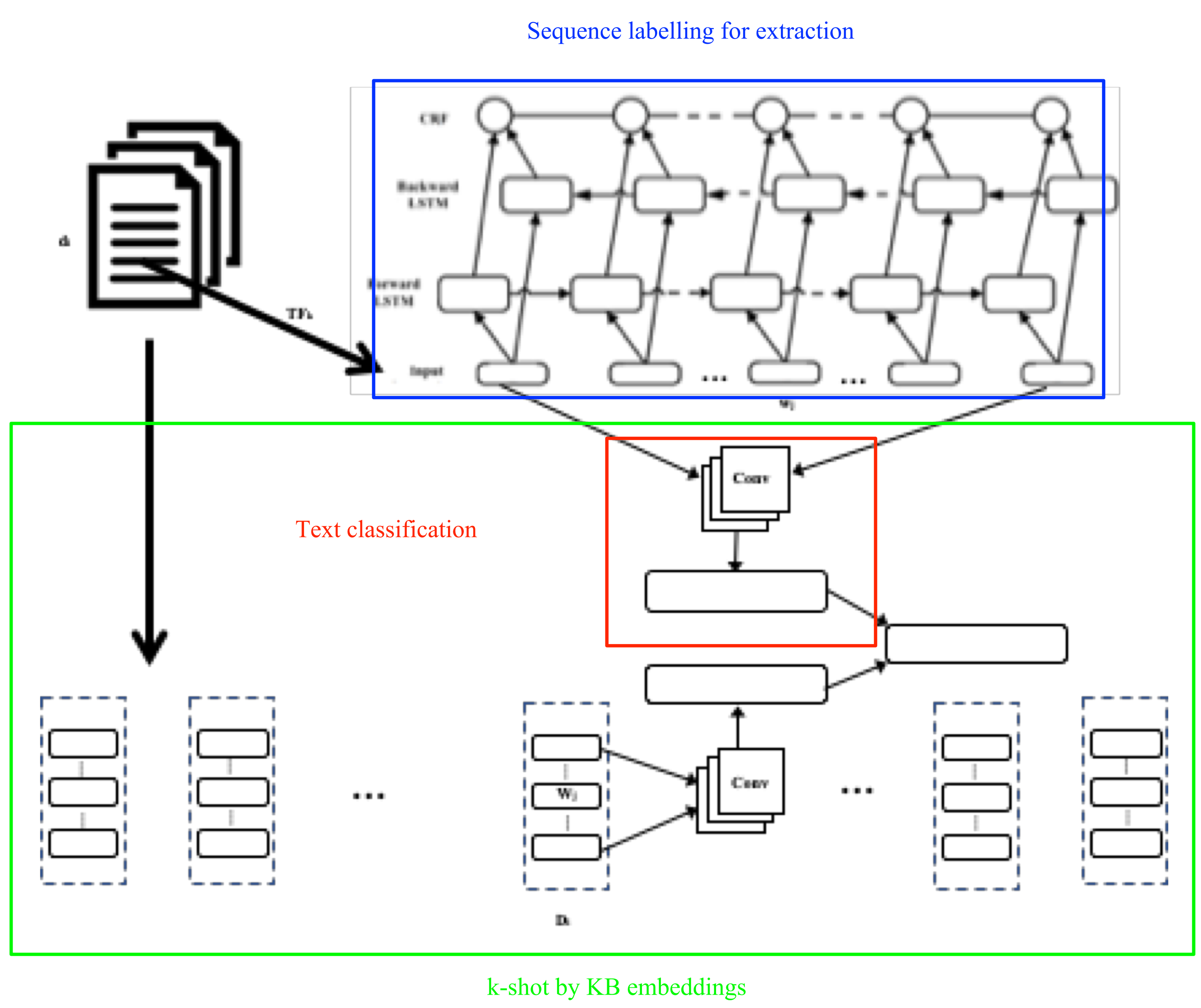
➢**Task**:
Identifying dataset surface forms (dataset mention extraction; **subtask A**) and associating the extracted mention to its referred dataset (dataset classification; **subtask B**).

Publication:
....Source: Monitoring the Future: National Survey on Drug Use, 1975-2009 …. Section 2 provides a brief summary of trends in adolescent drinking and smoking, using data for the US from the annual Monitoring the Future survey ….. Trends in Adolescent Drinking and Smoking: Monitoring the Future ….. Systematic annual data on the prevalence of underage drinking and smoking in the US are collected and tracked by several organizations. This section relies on data from the Monitoring the Future (MTF) ….

Datasets (Present): [...
56: Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, 1984;
101: Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, 1989; ...]

Datasets (Not Present): [...
100: Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 1996;
108: Current Population Survey, May 1973; ...]

❖ Models



➢**Shared Layer Extraction–Classification (‘SL E–C’)**
Share the output from the CNN-BiLSTM-Attn base, and substitute the CRF layer with a CNN layer for dataset classification.

➢**KB Shared Layer Extraction–Classification (‘KBSL E–C’)**
In this model, we leverage on the meta-information (description) of the dataset knowledge base to better support zero-shot learning

❖ Experiments & Results

➢**Scenarios:**
One-plus classification, zero-shot classification and zero-shot discovery

➢**Modeling Decisions:**
Size of the text-fragment
Sampling negative text-fragment

➢**Evaluation**:
On the **dev set**, the **test set** and on the **zero-shot test set.**

1) Randomly held out 7% of the datasets from the corpus and select the publications (219 documents in total) containing these datasets to form **zero-shot test set.**
2) Randomly hold out 225 publications to form the **test set**.
3) The **dev set** is split from the training set (5%) and has the same distribution and length as the training set.

	Partial			Exact		
Model	P	R	F1	P	R	F1
BiLSTM	29.4	32.1	30.7	11.2	12.8	12.0
CNN-BiLSTM	49.8	44.7	47.1	28.6	31.2	29.8
CNN-BiLSTM-CRF	54.1	44.8	48.9	35.6	33.8	34.7
CNN-BiLSTM-Attn-CRF	58.0	50.0	**53.7**	34.8	38.0	**36.4**
SL E-C	40.3	43.1	41.7	27.1	28.4	27.7

Performance of **test set** for mention extraction

Model	P	R	F1
BiLSTM	27.5	47.4	34.8
CNN	42.8	46.5	**44.6**
SL E-C	31.8	49.3	38.6

Performance of **test set** for classification

❖ Observations & Conclusion

We find CNN-BiLSTM-CRF and CNN models work best for dataset mention extraction and classification respectively.

Though appears to be a pipeline tasks strict pipeline configurations give poorer performance.

We identify that while mention extraction is primarily dependent on local signals the dataset classification uses a much wider context than just the mention.

Though the task appears to be easy for human turns out to be challenging for models due to extreme high output space and sparse per output signals.