How Does Beam Search Improve Span-Level Confidence Estimation in Generative Sequence Labeling?

Kazuma Hashimoto, Iftekhar Naim, and Karthik Raman

https://aclanthology.org/2024.uncertainlp-1.6/

UncertaiNLP at EACL 2024

Sequence Labeling / Text Segmentation

- Core NLP problem
  - NER
  - POS tagging
  - Slot filling
  - Search query understanding
  - etc...
- Uncertainty/confidence estimation
  - We may want to **drop** labeled spans with low confidence (to improve precision) [1]
  - An interactive system may want to **ask for confirmation** about labeled spans with low confidence [2]

x: [FIFA, World, Cup, 2022, in, Qatar],

y: [(FIFA, ASSOCIATION), (World Cup, EVENT), (2022, YEAR), (in, O), (Qatar, COUNTRY)],

Sequence Labeling with Text Generation

- It is a common strategy to select a (neural) model type for each task
  - Encoder+classifier
  - Encoder+decoder
  - Decoder (causal LM)
- These days, we are actively investigating the use of “**Generative AI**” for many tasks
  - Pretrained EncDec for research
    - BART, T5, …
  - Causal LMs for research/APIs
    - GPT, Gemini, Llama, …

x: [FIFA, World, Cup, 2022, in, Qatar],

y: [(FIFA, ASSOCIATION), (World Cup, EVENT), (2022, YEAR), (in, O), (Qatar, COUNTRY)],

$$y = \arg\max_{y'} p_\theta(y'|x)$$

Confidence Estimation for Sequence Labeling w/ Text Gen

- This work investigates various methods of confidence estimation for sequence labeling with text generation
    - Assumption
        - We use models that are fully controllable by us
            - We can have access to **token-level generation probability**
            - We can chose greedy search, **beam search**, random sampling, etc.
        - Targeted model
            - mT5 (or any similar fine-tunable models)
        - *Non* targeted models
            - API-only models

Google

Idea

- We use the best/top-1 prediction and would like to **estimate a confidence score of each span**
- In the example below, the labeled span "**in the area**" is wrong (and **inherently ambiguous**)
- The most straightforward approach
  - Span-level generation probability

$$c_\theta(y_i) = p_\theta(y_i|x, y_1, \ldots, y_{i-1})$$

- Can we have better ideas about how the model prefers the specific labeled span?
  - Let's look into the **statistics observed in other generated sequences**

Input	do you have listings of diners in the area
Gold	(do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, Location)
Top-5	1: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), $\boxed{\text{(in the area, Location)}}$
	2: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, Location)
	3: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, Location), (the, O), (area, Location)
	4: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, O), (in the area, Location)
	5: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, O)

Google

Proposal

- In which contexts does the labeled span appear?
    - Aggregated span probability (**AggSpan**)
        - Average span prob weighted by the context probs

$$c_\theta(y_i) = p_\theta(y_i|x) = \sum_z p_\theta(y_i|x, z)p_\theta(z|x)$$

- In which sequences does the labeled span appear?
    - Aggregated sequence probability (**AggSeq**)
        - Weighted count with the sequence-level probs

$$c_\theta(y_i) = \sum_{\hat{y}} p_\theta(\hat{y}|x)$$

Input	do you have listings of diners in the area
Gold	(do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, Location)
Top-5	1: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in the area, Location)
	2: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, Location)
	3: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, Location), (the, O), (area, Location)
	4: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, O), (in the area, Location)
	5: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, O)

Summary

- We have shown the effectiveness of the methods across 6 datasets
- An even better alternative of AggSeq is presented in our paper
    - Adjusting the beam size for each example dynamically
- Future work: how to work with API-only models?

Input	do you have listings of diners in the area
Gold	(do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, Location)
Top-5	1: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in the area, Location)
	2: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, Location)
	3: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, Location), (the, O), (area, Location)
	4: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, O), (in the area, Location)
	5: (do, O), (you, O), (have, O), (listings, O), (of, O), (diners, Cuisine), (in, O), (the, O), (area, O)
Span	(do, O)$_{\mathbf{0.99}}$, (you, O)$_{\mathbf{0.99}}$, (have, O)$_{\mathbf{0.99}}$, (listings, O)$_{\mathbf{0.99}}$, (of, O)$_{\mathbf{0.99}}$, (diners, Cuisine)$_{\mathbf{0.99}}$, (in the area, Location)$_{\mathbf{0.87}}$
AggSpan	(do, O)$_{\mathbf{0.99}}$, (you, O)$_{\mathbf{0.99}}$, (have, O)$_{\mathbf{0.99}}$, (listings, O)$_{\mathbf{0.99}}$, (of, O)$_{\mathbf{0.99}}$, (diners, Cuisine)$_{\mathbf{0.98}}$, (in the area, Location)$_{\mathbf{0.86}}$
AggSeq	(do, O)$_{\mathbf{1.0}}$, (you, O)$_{\mathbf{1.0}}$, (have, O)$_{\mathbf{1.0}}$, (listings, O)$_{\mathbf{1.0}}$, (of, O)$_{\mathbf{1.0}}$, (diners, Cuisine)$_{\mathbf{0.93}}$, (in the area, Location)$_{\mathbf{0.63}}$

References and Notes; Thank you for coming!

[1] https://arxiv.org/abs/2209.14694

[2] https://arxiv.org/abs/2203.12187

$$\frac{\sum_{z_{\mathcal{B}}} p_\theta(y_i|x, z_{\mathcal{B}}) p_\theta(z_{\mathcal{B}}|x)}{\sum_{z_{\mathcal{B}}} p_\theta(z_{\mathcal{B}}|x)}$$

$$\approx \frac{\sum_z p_\theta(y_i|x, z) p_\theta(z|x)}{\sum_z p_\theta(z|x)}$$

$$= \sum_z p_\theta(y_i|x, z) p_\theta(z|x),$$

$$\frac{\sum_{\hat{y}_{\mathcal{B}}} p(\hat{y}_{\mathcal{B}}|x)}{\sum_{j=1}^k p(y^{(j)}|x)} \approx \frac{\sum_{\hat{y}} p_\theta(\hat{y}|x)}{\sum_{y'} p_\theta(y'|x)}$$

$$= \sum_{\hat{y}} p_\theta(\hat{y}|x),$$