Neural Sequence-to-sequence Learning of Internal Word Structure

Tatiana Ruzsics, Tanja Samardžić
CorpusLab, URPP Language and Space, University of Zurich

Highlights

- Up to 4% improvement over a strong character-level sequence-to-sequence (cED) baseline for three languages
- Improvement over the previous state-of-the-art for two languages, while eliminating the need for external resources such as large dictionaries.
- Including corpus counts is beneficial to both encoder-decoder and classical statistical machine translation systems

Approach

Task: züricher → zürich|er

1. Combine cED with Language Model over morphemes (LM):

Assume corpus contains:	*Prediction:*	
züricherin → zürich|er|in	cED:	zürichen|er
	cED+LM:	zürich|er

2. Add Length Control (LC) as a difference in length between the input and its prediction:

Assume corpus contains:	*Prediction:*	
züricherin → zürich|er|in	cED:	zürichen|er
zürich → zürich	cED+LM:	zürich
	cED+LM+LC:	zürich|er

Tokens vs Types

	No. of	Error Rate (%)				
		Types Regime			Tokens Regime	
		cED	cSMT Baseline	cED Compar.	cED	cSMT Baseline
Total	24,606	0.19	0.18	0.23	0.16	**0.14**
Seen words	19,920	0.13	0.12	0.18	0.08	**0.07**
New comb.	3,959	0.44	**0.41**	0.42	0.47	**0.41**
New morph.	727	0.53	0.57	0.60	**0.48**	0.56

Table 2: Performance on the task of canonical segmentation for Chintang. Comparative setting for cED: training in types regime for the same number of iterations as in the individual setting of token regime.

References

[1] Ryan Cotterell, Tim Vieira, and Hinrich Schütze. A joint model of orthography and morphological segmentation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 664–669, 2016.

[2] Katharina Kann, Ryan Cotterell, and Hinrich Schütze. Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 961–967, 2016.

Morphological Segmentation Problem

Example 1 (Chintang)

```
a.              cuwa thaptakha
b.  cuwa   thapt   -a    -khag   -a
c.  water  move    -IMP  -see    -IMP
           across                [2sS]
d.    'Bring some water over here!'
```

Canonical segmentation:

thaptak**ha** → thapt | -a | -**khag** | -a

Decoding Algorithm

Beam Search: **expansion by morphemes** with combined score CS (weighted cED,LM,LC).

Input : Input word x, beam size n
Output: Predicted segmentation h

Initialize Hypotheses = ['<s>']
while not 1-best$_{CS}$(Hypotheses) is closed with '</s>': **do**
  New_Hypotheses=[]
  foreach $h_i \in$ Hypotheses **do**
    New_Hypotheses.append($h^1_{i+1}, \ldots, h^n_{i+1}$ from **SyncBeam$_{cED}$**(h_i))
  end
  Hypotheses = n-best$_{CS}$(New_Hypotheses)
end
return 1-best$_{CS}$(Hypotheses)

SyncBeam$_{cED}$: **expansion by characters** with cED score

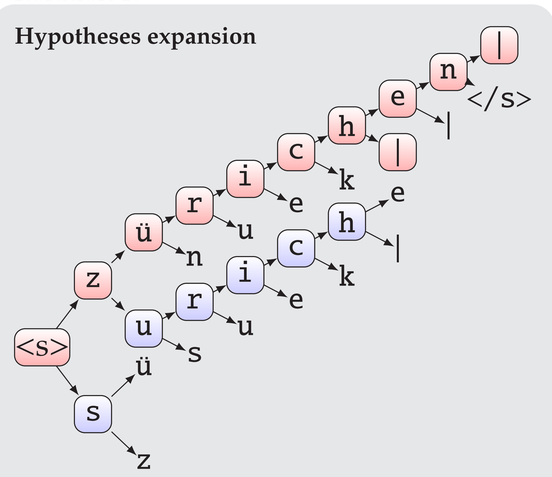
Input : Partial hypothesis h for an input word x, beam size n
Output: n-best expansions of h, closed with '|' or '</s>'.

Initialize Hypotheses=[h]
while not all h$_i \in$ Hypotheses are closed with '|' or '</s>': **do**
  New_Hypotheses=[]
  foreach $h_i \in$ Hypotheses **do**
    New_Hypotheses.append($h^1_{i+1}, \ldots, h^n_{i+1}$ from n-best expansions of h_i based on cED score)
  end
  Hypotheses =n-best$_{cED}$(New_Hypotheses)
end
return Hypotheses

Example

züricher → zürich | er

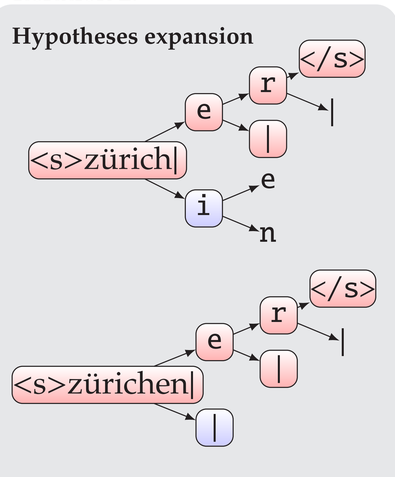
Iteration 1:



Scores Synchronization

Hypothesis	cED score	Combined score
zürichen|	**-1.5**	-5
zürich|	-8	**-2**

Iteration 2:



Scores Synchronization

Hypothesis	cED score	Combined score
zürich|er</s>	-8	**-3**
zürich|e|	-35	-9
zürichen|er</s>	**-1.5**	-6
zürichen|e|	-30	-12

Results

		Error Rate (%)				
		Types Regime				
		cED+LM	cED Baseline	cSMT Baseline	Joint*	cED+RR*
English	Total	**0.21** (.01)	0.22 (.01)	0.27 (.02)	0.27 (.02)	0.19 (.01)
	New comb.	0.15 (.03)	0.24 (.01)	-	-	-
	New morph.	0.23 (.01)	0.20 (.01)	-	-	-
German	Total	**0.19** (.00)	0.23 (.02)	0.24 (.02)	0.41 (.03)	0.20 (.01)
	New comb.	0.11 (.02)	0.33 (.03)	-	-	-
	New morph.	0.21 (.01)	0.20 (.01)	-	-	-
Indonesian	Total	**0.03** (.02)	0.07 (.01)	0.06 (.01)	0.10 (.01)	0.05 (.01)
	New comb.	0.02 (.03)	0.06 (.01)	-	-	-
	New morph.	0.09 (.03)	0.09 (.03)	-	-	-

Table 1: Performance on the task of canonical segmentation for English, German and Indonesian. cED+LM - character based encoder-decoder model fused with morpheme based language model. Baseline models: cED - character based encoder-decoder model, cSMT - character based statistical machine translation model. For reference only: Joint* - model of [1], cED+RR* - model of [2], not directly comparable since using external dictionary information.