

Semi-Supervised Learning for Neural Keyphrase Generation

EMNLP 2018

Task Definition

- Generate a list of keyphrases for the input document
 - Seq2Seq is the standard model for solving this task.
- Keyphrase sequence is created by connecting keyphrases with specific separator


Document:

In this paper, we consider an enthalpy formulation for a two-phase Stefan problem arising from the solidification of aluminum during **casting** process. We solve this free boundary problem in a time varying three-dimensional domain and consider convective heat transfer in the liquid phase. The resulting equations are discretized using a characteristics method in time and a **finite element** method in space, and we propose a numerical algorithm to solve the obtained nonlinear discretized problem. Finally, numerical results are given which are compared with industrial experimental measurements.

Keyphrase:

casting; **thermal**; **conduction**; **convection**; **finite element**

 in document

 not in document

Why Generation?

- Extraction-based Methods have been well studied.
 - Only the phrases appearing in the document will be extracted
- Keyphrase Generation is more challenging but more promising
 - Diverse and absent keyphrases can be discovered.

Document:

In this paper, we consider an enthalpy formulation for a two-phase Stefan problem arising from the solidification of aluminum during **casting** process. We solve this free boundary problem in a time varying three-dimensional domain and consider convective heat transfer in the liquid phase. The resulting equations are discretized using a characteristics method in time and a **finite element** method in space, and we propose a numerical algorithm to solve the obtained nonlinear discretized problem. Finally, numerical results are given which are compared with industrial experimental measurements.

Keyphrase:

casting; **thermal**; **conduction**; **convection**; **finite element**



in document



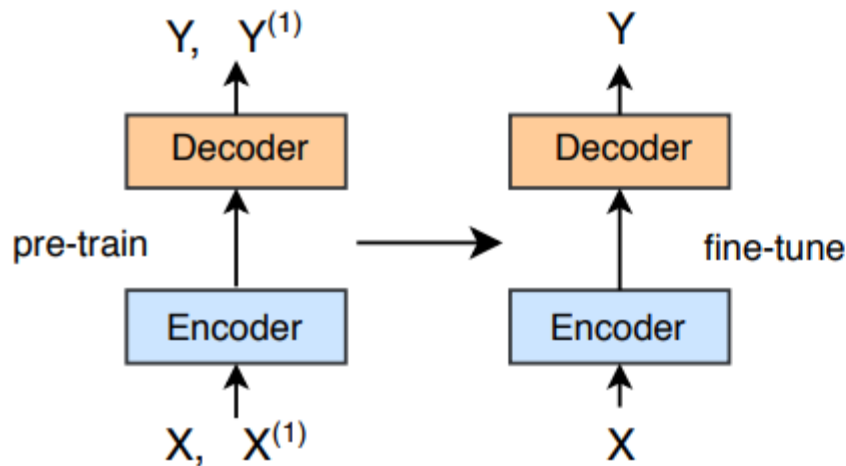
not in document

Motivation

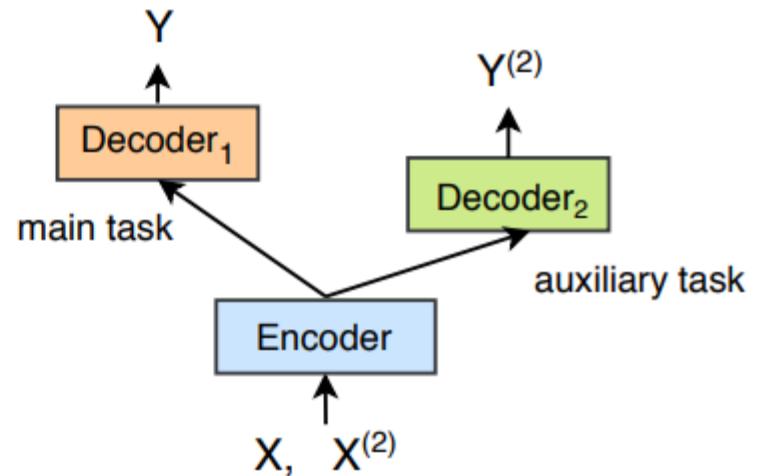
- The number of labeled training data for keyphrase generation is ****limited****.
- The knowledge hidden behind the unlabeled documents can help to generate keyphrases
 - Provide context information for keyphrase
 - Give some clues: e.g., keypphrases are likely to be NP or V
 - Improve model generalizability
- XXX

Model

- Base Model
 - Seq-to-Seq attentional model with Copy mechanism
- Two Semi-Supervised Learning Methods are proposed



(a) Synthetic Keyphrase Construction



(b) Multi-task Learning

Synthetic Keyphrase Construction

- Generate (synthesize) keyphrases for unlabeled documents
 - Unsupervised approaches (extraction-based) like TF-IDF and textrank
 - Self-learning algorithm (generation-based).
- Mix the **golden labeled data** and **synthetic labeled data** to pre-train the model
- Fine-tune model based on gold labeled data

Multi-task Learning

- Introduce auxiliary task and jointly train this task with the primary keyphrase generation task
 - In this paper, “title generation” is regarded as the auxiliary task
- The primary task correlates with the auxiliary task via parameter sharing
 - Share the same encoder network but have different decoders.
- xxx

Inference and Ranking

- Top-ranked keyphrase sequences are leveraged for producing the final keyphrases
 - Collect keyphrases from top-ranked beams to lower ranked beams
 - Keyphrases in the same sequence are ordered as in the generation process

Conclusion

- This paper provides a semi-supervised solution for neural keyphrase generation
 - Although the results are not exciting.
- Two methods for leveraging the large amount of unlabeled data
 - How to find the unlabeled data?
 - How to make use of the unlabeled data?
- Copy mechanism
 - The improvement maybe marginal on this task.