Survey of techniques for domain specific Named Entity Recognition

Manik Bhandari
Indian Institute of Science
manikb@iisc.ac.in

Abstract

This document is a brief summary of relevant work in the field of domain- specific Named Entity Recognition (NER).

1 Introduction

1.1 Define the problem

1.2 Major Challenges

1. Lack of corpus.

2. Lack of supervised NER tags - you have to manually define the set of entities that you are interested in so this set will always be small.

3. Disambiguation - same surface form can mean two things in different context.

2 Summary of Papers

2.1 AutoNER

AutoNER (Shang et al., 2018) has two contributions: Fuzzy LSTM CRF and Tie-or-Break scheme.

Fuzzy LSTM CRF Have to read about CRF and why they work.

They also introduce a training mechanism **Tie-or-Break Scheme**

3 Proposed Method

Syntactic-BERT solves all three major challenges.

1. Using transfer learning, we avoid the problem of lack of a huge corpus for getting contextualized embeddings.

2. We use additional low-level, syntactic tasks which provide low level supervision and force the model to learn syntactic information. This syntactic information like POS tags can be used by the model to predict Named Entities (Shang et al., 2018) [[Also cite POS paper]].

3. We use distant supervision paradigm and use *soft supervision* as described in AutoNER to allow our model to learn *soft labels*.

4 Experiments

5 Discussion

1. Preliminary experiments suggest that training BERT from scratch on a small corpus (20K sentences) is bad. This is a little surprising (but kind of expected) because the vocabulary size of this corpus is only around 10K.

2. Next in performance is Pretrained BERT without any language modeling fine tuning. Best so far (apart form SOTA) is BERT with language model fine tuning to domain specific corpus.

References

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064. Association for Computational Linguistics.

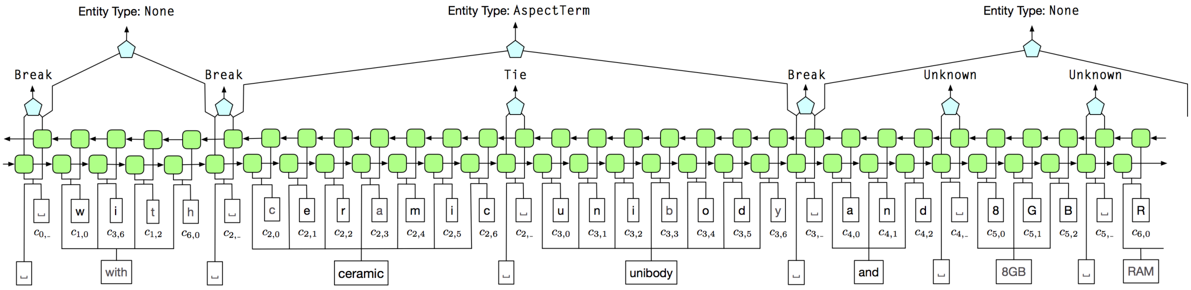

Figure 1: Overview of the Tie-or-Break scheme used in AutoNER.