Named Entity Generation using Sampling-based Structured Prediction

Guillaume Bouchard
Xerox Research Centre Europe
6 Chemin de Maupertuis
38240 Meylan, France
guillaume.bouchard@xerox.com

Abstract

The problem of Named Entity Generation is expressed as a conditional probability model over a structured domain. By defining a factor-graph model over the mentions of a text, we obtain a compact parameterization of what is learned using the SampleRank algorithm.

1 Introduction

This document describes the participation of the Xerox Research Centre Europe team in the GREC-NEG'10 challenge (http://www.nltg.brighton.ac.uk/research/gen chal10/grec/)

2 Model

Conditional random fields are conditional probability models that define a distribution over a complex output space. In the context of the Named-Entity Generation challenge, the output space is the set of possible referring expressions for all the possible mentions of the text. For example, assuming that we have the following text with holes (numbers are entity IDs):

> #1 was a Scottish mathematician,
> son of #2. #1 is most remembered
> as the inventor of logarithms
> and Napier's bones.

Then the possibilities associated with the entity #1 are:

1. John Napier of Merchistoun,
2. Napier,
3. he,
4. who,

and the possibilities associated with the entity #2 are:

1. Sir Archibald Napier of Merchiston,
2. he,
3. who.

Then, the output space is $Y = \{1,2,3,4\} \times \{1,2,3\} \times \{1,2,3,4\}$, representing all the possible combination of choices for the mentions. The solution $y = (1,1,3)$ corresponds to inserting the texts 'John Napier of Merchiston', 'Sir Archibald Napier of Merchiston' and 'he' in the holes of the text in the same order. This is the combination that is the closest to the original text, but a human could also consider that solution $y = (1,1,2)$ as being equally valid.

Denoting x the input, i.e. the text with the typed holes, the objective of the task is to find the combination $y \in Y$ that is as close as possible to natural texts.

We model the distribution of y given x by a factor graph: $p(y|x) \propto \prod_{c \in C} \phi_c(x,y)$, where C is the set of factors defined over the input and output variables. In this work, we considered 3 types of exponential potentials:

- Unary potentials defined on each individual output y_i. They include more than 100 features corresponding to the position of the mention in the sentence, the previous and next part of speech (POS), the syntactic category and funciton of the mention, the type and case of the corresponding referring expression, etc.
- Binary potentials over contiguous mentions include the distance between them, and the joint distribution of the types and cases.
- Binary potentials that are activated only between mentions and the previous time the

same entity was referred to by a name. The purpose of this is to reduce the use of pronouns referring to a person when the mentions are distant to each other.

To learn the parameter of the factor graph, we used the SampleRank algorithm (Wick et al., 2009) which casts the prediction problem as a stochastic search algorithms. During learning, an optimal ranking function is estimated.

3 Results

Using the evaluation software supplied by the GREC-NEG organizers, we obtained the folloing performances:

total slots	: 907
reg08 type matches	: 693
reg08 type accuracy	: 0.764057331863286
reg08 type matches including embedded	: 723
reg08 type precision	: 0.770788912579957
reg08 type recall	: 0.770788912579957
total peer REFs	: 938
total reference REFs	: 938
string matches	: 637
string accuracy	: 0.702315325248071
mean edit distance	: 0.724366041896362
mean normalised edit distance	: 0.279965348873838
BLEU 1 score	: 0.7206
BLEU 2 score	: 0.7685
BLEU 3 score	: 0.7702
BLEU 4 score	: 0.754
NIST score	: 5.1208

References

Michael Wick, Khashayar Rohanimanesh, Aron Culotta, and Andrew McCallum. 2009. SampleRank: Learning preferences from atomic gradients. *Neural Information Processing Systems (NIPS) Workshop on Advances in Ranking*.