Neural coreference resolution with limited lexical context and explicit mention detection for oral French: supplemental material

Loïc Grobol
Lattice CNRS, 1 rue Maurice Arnoux, 92120 Montrouge, France
ALMAnaCH, Inria, 2 rue Simone Iff, 75589 Paris, France
`loic.grobol@inria.fr`

1 Hyperparameters

In this section, we give the details of the specific hyperparameters of our system that are not fundamental to he principle of our architecture, but that might be useful for reproduction or comparisons to future works.

Words representations We use character embeddings of dimension 50, and encode them using network with a BiGRU hidden layer of dimension 150, the final state of which is passed through a 50-dimensional feedforward layer to get the final character-level representation of each word.

Spans encoding The layers of the span encoder, recurrent or not, are all 300-dimensional, except for the output layer $\mathrm{FFNN}_{\mathrm{out}}$, which is 500-dimensional (so the final span embeddings are of dimension 500) and the feedforward parts (attention all have a single hidden layer. The “span length” feature is first digitized through the same buckets as the mention distances in Lee et al. (2017) and projected to 20-dimensional embeddings.

Mention detection The mention detector module is a feedforward neural network network with a single hidden layer of dimension 150

Antecedent finding We set the cutoff for antecedent candidate pruning after the coarse scoring step to the 25 candidates with the highest scores. In addition to their respective span embeddings, we also use the following features in the mention-antecedent pair scoring module: distance in words, mentions and utterances, speaker agreement and overlap. The last two are boolean features and the distances are bucketed, again using the same buckets as Lee et al. (2017), and projected to 20-dimensional embeddings.

Regularization layers All the layers were subject to Dropout during training, with probability 0.6 for the word representations, 0.4 for the antecedent scoring layers, 0.3 for the span encoding layers and 0.2 for all of the other layers. We also apply Layer Normalization (J. L. Ba et al. 2016) on the final output of the span encoding module, purely to ease our monitoring of its outputs during training as it did not seem to have any significant impact on the training process or the final performances.

Finally, we use the leaky ReLU (Maas et al. 2013) non-linearity instead of the original ReLU, as it helped the network to move out from the local optimum of always predicting the “None” class during the mention detection training.

Training The network was trained for both tasks with early stopping subject to its performances on the development sets, which topped under 10 epochs in all of our experiments.

For mention detection, we use a base learning rate of 10^{-3}, with a linear warmup over the first 1000 mini-batches and decayed by a factor of 0.7 after each subsequent epoch. We also apply a weight decay of 10^{-5} to all the trainable parameters. For antecedent scoring we use a constant learning rate 10^{-4} and no weight decay. All the other optimizer parameters are the defaults used in the original Adam implementation (Kingma and J. Ba 2014).

Finally, we use mini-batches of size 30 for mention detection and 10 for antecedent scoring and shuffle the train sets after each epochs to ensure their homogeneity.

2 Dataset partition

The detailed list of the documents of the ANCOR corpus and their respective subcorpora in our training/development/test partition is available in the attached `ancor.json` file.

As mentionned in the main material, we tried to stay close to Désoyer et al. (2015) with about 60 % of the corpus devoted to the training set. However, we chose to keep most of the rest to the test set, in order to provide more significant final scores.

The final distribution in 59 %/12 %/29 %, with a fairly homogeneous distribution of the different subcorpora, in order to minimize the disparities caused by their various levels of interactivity and topics.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *NeurIPS 2016 Deep Learning Symposium*, 2016: arXiv: `1607.06450`.

Adèle Désoyer, Frédéric Landragin, Isabelle Tellier, Anaïs Lefeuvre, and Jean-Yves Antoine. 2015. Coreference Resolution for Oral Corpus: a machine learning experiment with ANCOR corpus. *Traitement Automatique des Langues*. Traitement automatique du langage parlé, 55.2, May 2015: 97–121.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*. San Diego, California. arXiv: `1412.6980`.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-End Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197. Association for Computational Linguistics. København, Danmark.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.