Topics for Projects

Giuseppe Attardi

Human Language Technologies

Dipartimento di Informatica

Università di Pisa

Challenges

- COVID-19 Global Hackathon
  - https://covid-global-hackathon.devpost.com/
- BioASQ (http://bioasq.org)
- Pharma CoNER (http://temu.bsc.es/pharmaconer/)
- Loop Q Prize (https://www.loopqprize.ai)
- The Conversational Intelligence Challenge 2 (http://convai.io)

Question Generation

- INVALSI (https://www.invalsi.it/invalsi/)
  - Question generation from Wikipedia articles

Chatbot

- Alexa Topical Chat Dataset
    - https://github.com/alexa/alexa-prize-topical-chat-dataset
    - Identify transitions between topics
    - Suggest sources of information
- The Conversational Intelligence Challenge 2 (ConvAI2)
  convai.io/

IWPT Shared Task

- The Enhanced Universal Dependency Shared Task at IWPT 2020 involves dependency parsing from plain text.
- This involves several subtasks:
  - Tokenization using DL
  - POS using DL
  - Morphological analysis
  - Depenedency parsing
  - Enhanced dependencies
- Timeline:
  - Test data: April 2, 2020
  - Submission: **April 22, 2020**

CoNLL 2018 UD Parsing

- Parsing Universal Dependencies for the CoNLL 2018 Shared Task:
    - Experiment “Left-to-right dependency parsing with pointer network” https://arxiv.org/pdf/1903.08445.pdf

CoNLL 2018: Deep Learning Tokenizer

- CoNLL 2018 challenge requires a tokenizer for all the Universal Dependency TreeBanks
- Build a DL tokenizer using Keras based on the approach of:
  - Basile, Valerio and Bos, Johan and Evang, Kilian *A General-Purpose Machine Learning Method for Tokenization and Sentence Boundary Detection* (2013), http://gmb.let.rug.nl/elephant/

CoNLL 2018: Deep Learning POS

- Depling 2016 challenge requires tokenizer for any of the Universal Dependency TreeBank
- Build a DL POS using CNN, for example a LSTM that uses word embeddings and possible charcater embeddings.

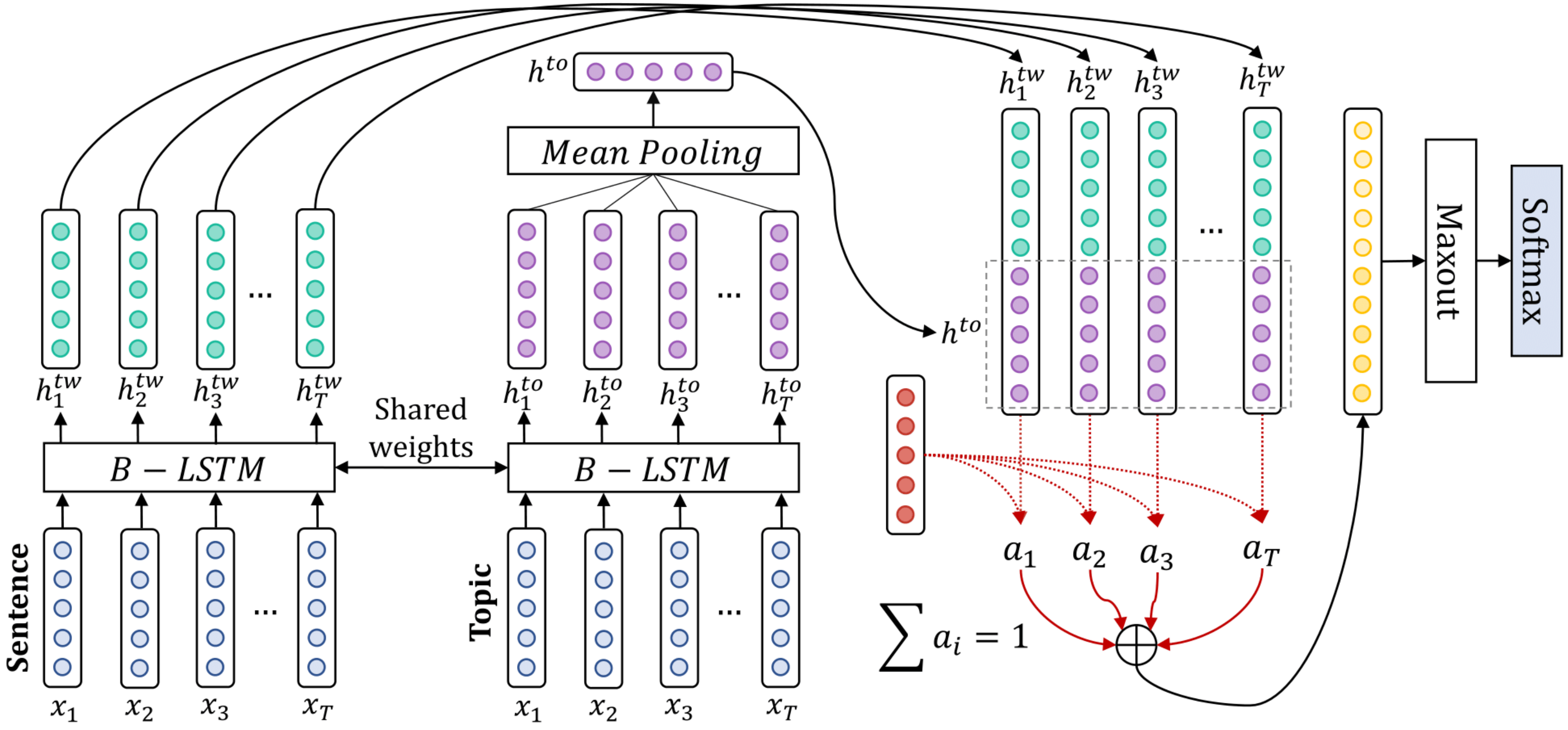
CoNLL 2018: Deep Learning Morph Analyzer

- CoNLL 2018 challenge requires dealing with all the Universal Dependency TreeBanks
- Build a DL morphological analyzer that computes morphological embeddings for each word, using Keras and character embeddings.

Evalita 2016-2018

- www.evalita.it/2016
  - POSTWITA
  - QA4FAQ
  - NEEL-IT
- www.evalita.it/2018
  - ABSITA
  - HaSpeeDe
  - NLP4FUN (more statistics than linguistics?)
  - Timeline
    - Data Release: May 28, 2018
    - Evalutation: September 10-16, 2018

Possible Approach for ABSITA



A Siamese Bidirectional LSTM with context-aware attention.

- Baziotis et al. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. www.aclweb.org/anthology/S17-2126
- Code: https://github.com/cbaziotis/datastories-semeval2017-task4

Question Answering Tasks

- Tensorflow 2.0 QA
  - https://www.kaggle.com/c/tensorflow2-question-answering
- SemEval 2017

  Task 3
- Evalita 2016

  QA4FAQ
- SQuAD

  https://towardsdatascience.com/nlp-building-a-question-answering-model-ed0529a68c54
- Movie QA

  http://movieqa.cs.toronto.edu/home/
- Natural Language Interfaces for Web of Data (NLIWoD4)

  http://2018.nliwod.org/challenge

Chatbots

- AWS Chatbot Challenge
  - https://aws.amazon.com/events/chatbot-challenge/
- Ubuntu Dialog Corpus:
  - https://github.com/rkadlec/ubuntu-ranking-dataset-creator

Neural Machine Translation

- English-Italian
  - Europarl Corpus
  - Ses2Seq TensorFlow Tutorial
- References:
  - D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate. http://arxiv.org/pdf/1409.0473v6
  - Zhang, X., & LeCun, Y. (2015). Text Understanding from Scratch. http://arxiv.org/abs/1502.01710

Twitter

- Modeling Political Bias
    - Use Italian Tweets collection
- Detecting Toxic Comments
    - Use Italian Tweets collection and Evalita 2018 HaSpeeDe corpus

Deep Learning for Sentiment Analysis

- Annotated Data: SemEval training set
  - http://alt.qcri.org/semeval2017/task4/index.php?id=data-and-tools
- Unannotated Data: 50 million tweets
- CNN approach:
  - Code: DeepNL, https://github.com/attardi/deepnl
  - Article: A. Severyn, A. Moschitti.*UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification*
- BiLSTM approach:
  - Baziotis et al. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. www.aclweb.org/anthology/S17-2126
  - Code: https://github.com/cbaziotis/datastories-semeval2017-task4

POS tagging using Word Embeddings

- Data: Evalita 2016
- Embeddings: http://tanl.di.unipi.it/embeddings/
- Article: Stratos, M. Collins. Simple Semi-Supervised POS Tagging. http://www.cs.columbia.edu/~stratos/research/naacl15semipos.pdf

Medical texts

- Predicting side effects of drugs
  - Using collection of Italian medical record on kidney and heart diseases
- Negation/Speculative Scope Detection
  - BioScope Corpus: http://rgai.inf.u-szeged.hu/index.php?page=bioscope
- Semantic QA on medical texts:
  - BioASQ datasets: bioasq.org/

Negation/Speculation Scope

- Determine the scope of negative or speculative statements:
  - The lyso-platelet had no effect
  - MnlI-AluI could suppress the basal-level activity
- Approach:
  - Classifier for identifying cues
  - Classifier to determine scope
- Data
  - BioScope collection

Relation Extraction

- Exploit word embeddings as features + extra hand-coded features
- SemEval 2014 Relation Extraction dataset

Fake News Detection

- Stance Detection dataset for FNC-1
  - http://www.fakenewschallenge.org