Natural Language Processing (Almost) From Scratch

Collobert, Weston, Bottou, Karlen, Kavukcuoglu and Kuksa

Overview

1. Introduction
   - Overview
   - Motivation
   - Tasks
2. Approach
   - Window based
   - Sentence based
3. Evaluation
   - Supervised
   - Unsupervised Model
   - Results
4. Summary

Motivation

- Aim is to create a model that is capable of performing well on several NLP tasks.
- Initially ignore most linguistic knowledge.
- Minimise amount of pre-processing – allow the model to create features.

Tasks

- Four tasks chosen were:
  - Part of Speech Tagging
  - Chunking
  - Named Entity Recognition
  - Semantic Role Labelling
- All of these tasks can be thought of as mapping words to tags.
- Benchmark systems chosen:
  - POS Tagging: Toutanova et al. (2003)
  - Chunking: Sha and Pereira (2003)
  - NER: Ando and Zhang (2005)
  - SRL: Koomen at al. (2005)

Multi-layer Neural Networks

- A multi-layer neural network can be thought of as a series of functions.
- Including non-linear layers, such as the hard tanh function, allows more complex features to be modelled.
- Trained by backpropagation.

Feature Table

- Words are converted into vectors of real numbers by the lookup table layer.
- The length of these vectors are a parameter of the model.
- The values of the vector are set during training.
- Features other than words can also be represented in this manner.

Window-based Model

- The window based version of the model attempts to determine the tag of a word based on a T word window around the word of interest.
- T is a parameter of the model.
- After converting the words to a matrix of representation vectors, the columns are concatenated.
- This vector is then passed into a linear layer,
- then a hard tanh layer to introduce non-linearity,
- then a final linear layer, which is task-dependent, and has as many outputs as there are tags for that task.

Sentence-based Model

- SRL needs to look at the whole sentence due to interactions between distant words.
- Slightly more complicated, as the length of sentences can vary but the length of the representation vector needs to remain constant.
- Requires a convolutional neural network.
- Since the word of interest is no longer always the one in the centre, another feature is required to represent this.

Sentence-based Model

- After converting the words to a matrix of representation vectors, instead of concatenating, use a convolution layer.
- This convolution layer creates a number of representation vectors by using a sliding window along the sentence.
- A ‘max’ layer then selects the maximum value from each row.
- This representation is then passed into the same last three layers as for the window-based model.

Supervised Model

- Initially, both models were trained in a supervised manner.
- Dictionary consisted of 100,000 most common words in WSJ.
- pre-processing: lowercase, capitalisation encoded as feature, numbers replaced with 'NUMBER' keyword.
- This gave performance slightly worse than the benchmark systems.

Supervised Model

FRANCE 454	JESUS 1973	XBOX 6909	REDDISH 11724	SCRATCHED 29869	MEGABITS 87025
PERSUADE	THICKETS	DECADENT	WIDESCREEN	ODD	PPA
FAW	SAVARY	DIVO	ANTICA	ANCHIETA	UDDIN
BLACKSTOCK	SYMPATHETIC	VERUS	SHABBY	EMIGRATION	BIOLOGICALLY
GIORGI	JFK	OXIDE	AWE	MARKING	KAYAK
SHAHEED	KHWARAZM	URBINA	THUD	HEUER	MCLARENS
RUMELIA	STATIONERY	EPOS	OCCUPANT	SAMBHAJI	GLADWIN
PLANUM	ILIAS	EGLINTON	REVISED	WORSHIPPERS	CENTRALLY
GOA'ULD	GSNUMBER	EDGING	LEAVENED	RITSUKO	INDONESIA
COLLATION	OPERATOR	FRG	PANDIONIDAE	LIFELESS	MONEO
BACHA	W.J.	NAMSOS	SHIRT	MAHAN	NILGIRIS

- As the goal is to learn a generalizable model, it is useful to look at the word representations.
- Ideally, words with similar meanings would have similar representations.

Unsupervised Model

- Trained language models on large amounts of unlabelled data.
- The language models used the same structure as the window based network described above.
- Uses a ranking criterion to score the window.
  - LM1 was trained on 631 million words from Wikipedia, pre-processing was the same, dictionary: 100,000 most common words from WSJ.
  - LM2 was trained on the Wikipedia dataset, as well as 221 million words from Reuters, pre-processing was the same, dictionary: 100,000 most common words from WSJ, and 30,000 most common words from Reuters.

Unsupervised Model

FRANCE 454	JESUS 1973	XBOX 6909	REDDISH 11724	SCRATCHED 29869	MEGABITS 87025
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

- Initialising the supervised model with the word representations learnt by the language model improves the final word representations.

Results

Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
NN+WLL	96.31	89.13	79.53	55.40
NN+SLL	96.37	90.33	81.47	70.99
NN+WLL+LM1	97.05	91.91	85.68	58.18
NN+SLL+LM1	97.10	93.65	87.58	73.84
NN+WLL+LM2	97.14	92.04	86.96	58.34
NN+SLL+LM2	97.20	93.63	88.67	74.15

Results

Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL
Benchmark Systems	97.24	94.29	89.31	77.92
NN+SLL+LM2	97.20	93.63	88.67	74.15
NN+SLL+LM2+Suffix2	97.29	–	–	–
NN+SLL+LM2+Gazetteer	–	–	89.59	–
NN+SLL+LM2+POS	–	94.32	88.67	–
NN+SLL+LM2+CHUNK	–	–	–	74.72

Summary

- Aim:
  - Create a general purpose model for NLP tasks.
  - Avoid using linguistic knowledge to create features.
- The model performs close to state-of-the-art for all four tasks.
- Using linguistic knowledge to build extra features for the model further improves performance.
- The model is quicker and uses less memory than the state-of-the-art systems, as it doesn't need to calculate and store many complex features.