THESIS PROPOSAL

LEARNING REPRESENTATIONS
FOR TEXT-LEVEL DISCOURSE PARSING

Copyright © 2015 gw0 [http://gw.tnode.com/] <gw.2015@tnode.com>
OVERVIEW

• motivation
• discourse parsing
  ▪ PDTB-style
• deep learning architectures
  ▪ sequence processing
  ▪ word embeddings
• our approach
  ▪ key ideas
  ▪ guided layer-wise multi-task learning
• progress
MOTIVATION

• natural language processing (NLP)
  ▪ large pipelines of independently-constructed components or subtasks
  ▪ traditionally hand-engineered sparse features based on language/domain/task specific knowledge
  ▪ still room for improvement on challenging NLP tasks

• deep learning architectures
  ▪ backpropagation could be the one learning algorithm to unify learning of all components
  ▪ latent features/representations are automatically learned as distributed dense vectors
  ▪ surprising results for a number of NLP tasks
**DISCOURSE PARSING**

- **discourse**: a piece of text meant to communicate specific information (clauses, sentences, or even paragraphs)
- understood only in relation to other discourses, their joint meaning is larger than individual unit's meaning alone

[**Index arbitrage doesn't work**]_{arg1},

*and* [**it scares natural buyers of stock**]_{arg2}.


[**But**]_{arg2}

*if* [**this prompts others to consider the same thing**]_{arg1},

*then* [**it may become much more important**]_{arg2}.

He added [that "having just one firm do this isn't going to mean a hill of beans"]argin.

But [if this prompts others to consider the same thing, then it may become much more important"]argout.

— PDTB-style, id: 14904, type: explicit, sense: Comparison.Concession

In addition, Black & Decker had said it would sell two other undisclosed Emhart operations if it received the right price. [Bostic is one of the previously unnamed units, and the first of the five to be sold]argin

[The company is still negotiating the sales of the other four units and expects to announce agreements by the end of the year]argin. [The five units generated sales of about $1.3 billion in 1988, almost half of Emhart's $2.3 billion revenue]argout. Bostic posted 1988 sales of $255 million.

— PDTB-style, id: 12886, type: entrel, sense: EntRel
PDTB-STYLE DISCOURSE PARSING

- **Penn Discourse Treebank** adopts the predicate-argument view and independence of discourse relations
  - 2159 articles from the Wall Street Journal
  - 4 discourse sense classes, 16 types, 23 subtypes
- also called shallow discourse parsing
  - discourse relations are not connected to each another to form a connected structure (tree or graph)
  - adjacent/non-adjacent units in same/different sentences
- primary goals
  - locate explicit or implicit discourse **connective**
  - locate text spans for **argument 1 and 2**
  - predict **sense** that characterizes the nature of the relation
DEEP LEARNING ARCHITECTURES

- multiple layers of learning blocks stacked on each other
- beginning with raw data, its representation is transformed into increasingly higher and more abstract forms in each layer, until final low-dimensional features for a given task
SEQUENCE PROCESSING

Text documents of different lengths are usually treated as a sequence of words:

- transition-based processing mechanisms
- recurrent neural networks (RNNs)
  - applying the same set of weights over the sequence (temporal dimension) or structure (tree-based)
WORD EMBEDDINGS

Represent text as numeric vectors of fixed size:

- **word embeddings**: SGNS (word2vec), GloVe, ...
- feature/phrase/document embeddings
- character-level convolutional networks

**Unsupervised** pre-training helps develop natural abstractions. Sharing word embedding in **multi-task learning** improves their performance in the absence of hand-engineered features.
OUR APPROACH

• PDTB-style end-to-end discourse parser
• one deep learning architecture instead of multiple independently-constructed components
• almost without any hand-engineered NLP knowledge

Input:
• tokenized text documents (from CoNLL 2015 shared task)

Output:
• extracted PDTB-style discourse relations
  ▪ connectives
  ▪ arguments 1 and 2
  ▪ discourse senses
KEY IDEAS

• unified end-to-end architecture
  ▪ backpropagation as the one learning algorithm for all discourse parsing subtasks and related NLP tasks

• automatic learning of representations
  ▪ in hidden layers of deep learning architectures (bidirectional deep RNN/LSTM)

• shared intermediate representations
  ▪ partially stacked on top of each other to benefit from each others representations

• guided layer-wise multi-task learning
  ▪ jointly learning all discourse parsing subtasks and related NLP tasks including unsupervised pre-training
GUIDED LAYER-WISE MULTI-TASK LEARNING

Shared intermediate representations

Multi-task learning

discourse sense classification subtask

increasingly more difficult tasks

argument boundary detection subtask

discourse presence identification subtask

part-of-speech tagging

unsupervised language model

word embeddings (lowest representation)

input

time

text

some

input
PROGRESS

- technology
  - Python
  - Theano: fast tensor manipulation library
  - Keras: modular neural network library

- resources and inputs
  - pre-trained word2vec lookup table (on Google News)
  - tokenized text documents as input
  - POS tags of input tokens

- evaluation (from CoNLL 2015 shared task)
  - performance in terms of precision/recall/F1-score
  - explicit connectives, argument 1, 2 and combined extraction, sense classification, overall
COMPLICATION OR USEFUL?

Experiments with single-task learning with bidirectional deep RNN for discourse sense tagging:

input: ... cool it because they 're ...

vector:

\[
\begin{bmatrix}
6 \\
7
\end{bmatrix}, \begin{bmatrix}
5 \\
0
\end{bmatrix}, \begin{bmatrix}
0 \\
2
\end{bmatrix}, \begin{bmatrix}
4 \\
2
\end{bmatrix}, \begin{bmatrix}
6 \\
0
\end{bmatrix}
\]

h1:

\[
\begin{bmatrix}
1 \\
0 \\
0 \\
0
\end{bmatrix}
\]

h2:

\[
\begin{bmatrix}
1 \\
0 \\
0 \\
0
\end{bmatrix}
\]

h3:

\[
\begin{bmatrix}
0 \\
0 \\
0 \\
0
\end{bmatrix}
\]

prediction label:

arg1, arg1, conn, arg2, arg2
SINGLE-TASK RESULTS

- long training time for randomly initialized weights
  - lower tasks improve initialization
- overfitting training data
  - more tasks improve generalization

FUTURE EXPERIMENTS

- various discourse parsing subtasks
- various related NLP tasks (chunking, POS, NER, SRL, ...)
- different representation structures
- different activation, optimization, architectures
- long short-term memory (LSTM)
- neural Turing machines (NTM)
DOES IT MAKE SENSE?

I would like to hear your feedback and ideas for my thesis proposal.

THANK YOU

http://gw.tnode.com/deep-learning/acl2015-presentation/

Copyright © 2015 gw0 [http://gw.tnode.com/] <gw.2015@tnode.com>