Research Background and Perspectives

Erion Çano

Institute of Formal and Applied Linguistics

Faculty of Mathematics and Physics

Charles University, Czech Republic

Definition and Applications of SA

Sentiment Analysis

Computational examination of sentiments, opinions, and attitudes expressed in text from an opinion holder towards an entity.

Sentiment Classification

Determining the polarity of an opinion in a text unit about an entity. It can be document-level, sentence-level or aspect-level.

Applications

Market surveys and predicitons, brand/product popularity analysis, client/product profiling, political surveys, counter-terrorism, etc.

Sentiment Classification Tasks

Sentiment Classification (SC)

The task of determining the polarity of an opinion about an entity.

Document-level SC

Performing SC task on a document of a single opinion holder about a single entity.

Sentence-level SC

Performing SC task on subjective sentences of one opinion each, from a single opinion holder.

Aspect-level SC

Performing SC task on different and specific aspects of an entity.

SA Techniques

SA techniques can be categorized as:

- **Supervised Learning**
  - Considering SA as *Document Classification*
  - Pang and Lee, 2002-2005
- **Semantic orientation and Lexicons**
  - PMI-IR algorithm of Turney, 2001
  - WordNet-Affect lexicon
- **Hybrid**
  - Using lexicons to create training texts
  - Seed words extended with synonyms

Text Preprocessing Steps

- Cleaning and Tokenization.

  - Removed remaining html tags.

  - Kept in smiley symbols like :-), :), :(, :-(, :P, :D

- Part of Stopwords removed.

  - Cleared useless terms like "the", "that", "by".

  - Retained negation residues like "don", "didn", "hasn".

- Clipping and Padding.

  - Clipped few very long documents.

  - Zero-padded shorter documents.

Word Representations: Bag of Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Severe sparsity and dimensionality issues on large V...☹

Word Representations: Word Embeddings

"*You shall know a word by the company it keeps.*"

– J. R. Firth, 1957

Word Embeddings

- ☺ Dense and low-dimensional
- ☺ Complexity scales linearly w.t.r V
- ☺ Preserve word order in phrases
- ☺ Capture semantic and syntactic similarities
- ☹ Require big text corpora to train
- ☹ Computationally expensive to train

Representation of Words

Matrix representaton of “your shirt looks nice”:

d = 5

your	0.23	0.18	0.34	0.76	0.62
shirt	0.64	0.23	0.21	0.03	0.83
looks	0.98	0.59	0.76	0.65	0.45
nice	0.11	0.43	0.30	0.22	0.92

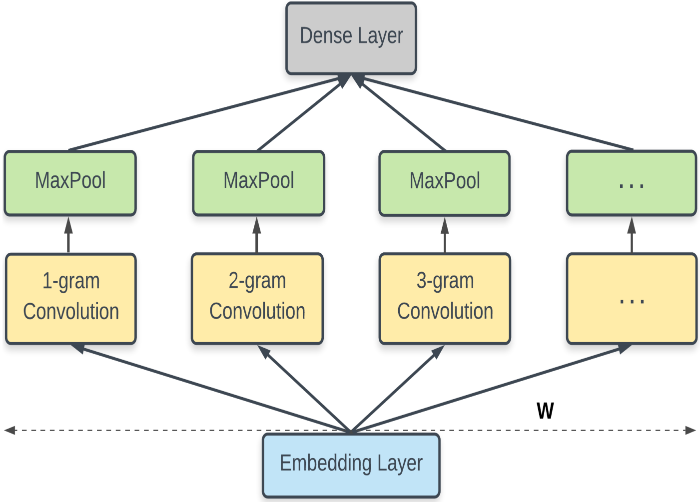
Embeddings sourced from pretrained GoogleNews collection.

Experimental Datasets

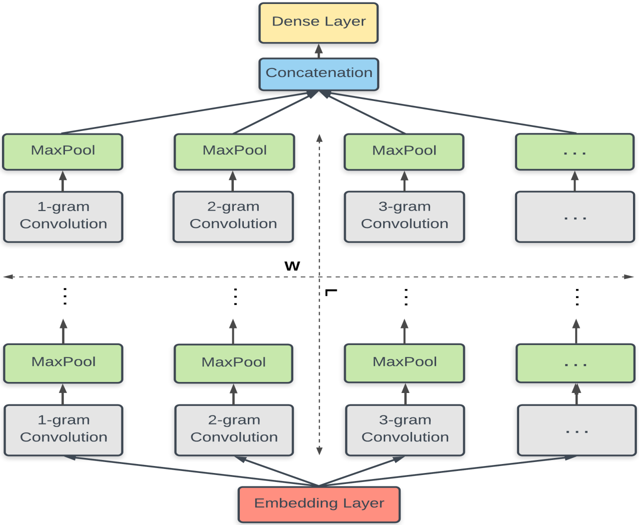
Dataset	Docs	MinL	AvgL	MaxL	UsedL
Mlpn (song lyrics)	5K	23	227	2733	450
Sent (Sentences)	10K	1	17	46	30
Imdb (movie reviews)	50K	5	204	2174	400
Phon (phone reviews)	232K	3	47	4607	100
Yelp (yelp reviews)	598K	1	122	963	270

- Different domain tasks and data types
- Both small (Mlpn) and big (Yelp) datasets
- Both long (Imdb) and short (Phon) documents

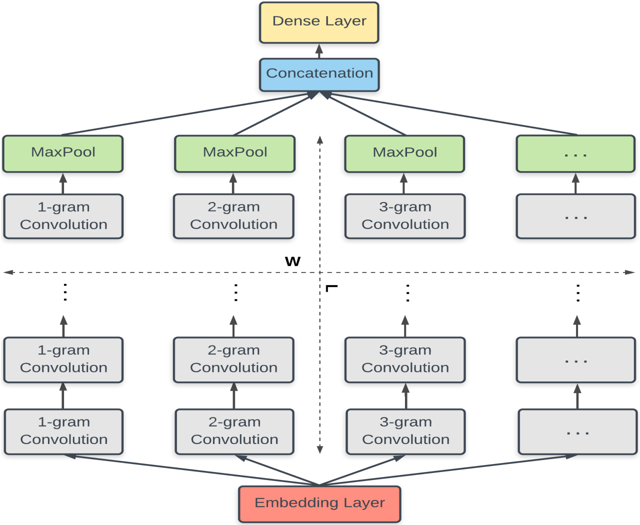
Multi-Channel Network Structures



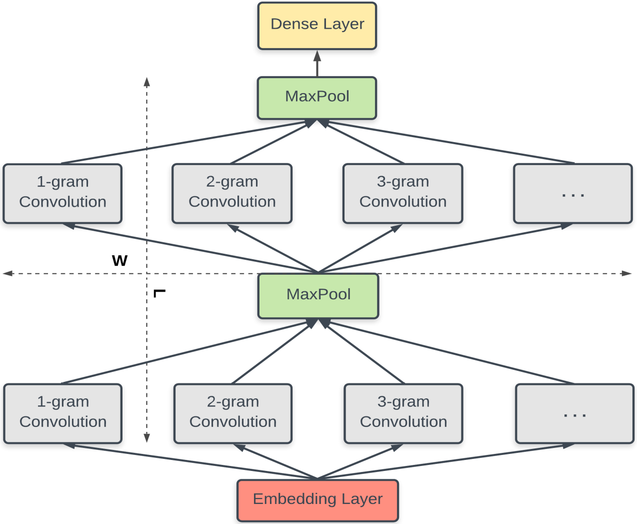
NgramCNN Basic Architecture



NgramCNN Pyramid Architecture



NgramCNN Fluctuating Architecture



Baseline Models

- Single LSTM
- Single Convolution-Pooling
- Bidirectional LSTM with max-pooling
- Bidirectional LSTM with Convolution-Pooling
- Logistic Regression with tf-idf
- Support Vector Machine with tf-idf

Comparative Accuracy Scores

Network	Sent	Imdb	Phon	Yelp
NgCNN Basic	79.87	90.77	**95.92**	**94.88**
NgCNN Pyramid	79.52	**91.21**	95.70	94.83
NgCNN Fluctuate	77.41	89.32	93.45	92.27
Optimized LR	81.63	89.48	92.46	91.75
Optimized SVM	82.06	88.53	92.67	92.36
SingleCNN	81.79	89.84	94.25	93.86
SingleLSTM	80.33	84.93	93.71	90.22
BLSTM-POOL	80.96	85.54	94.33	91.19
BLSTM-2DCNN	**82.32**	85.70	95.52	91.48

Further Observations

- *Deep feature networks with simple classifiers are top performers on texts, same as on images.*
- *Basic NgramCNN architecture is fast and highly accurate on long documents.*
- *LSTM-based architectures are slower to train and perform poorly on long documents.*
- *For small datasets, traditional linear classifiers like LR or SVM could be good enough.*

[4] E. Çano, M. Morisio: A Deep Learning Architecture for Sentiment Analysis

More about me...

Old Website: http://softeng.polito.it/erion

New Website: https://ufal.mff.cuni.cz/erion-cano

Definition of Text Summarization

Text Summarization (TS)

Distilling the most important information in a text to produce an abridged version.

Types of TS

- Single-document vs Multi-document
- Extractive vs Abstractive
- Generic vs Query-driven
- Informative vs Indicative

Why to Summarize...?

- Simplify and abbreviate text (Abstracts)
- Summary of email threads (Subjects)
- Action items from a meeting (Discussions)
- Generating news of an event (Stories)
- Basic opinions about an item (Reviews)
- Answering user questions (Queries)

Extractive vs Abstractive

Extractive TS

The generated summary is a selection of relevant sentences from the source text in a copy-paste fashion.

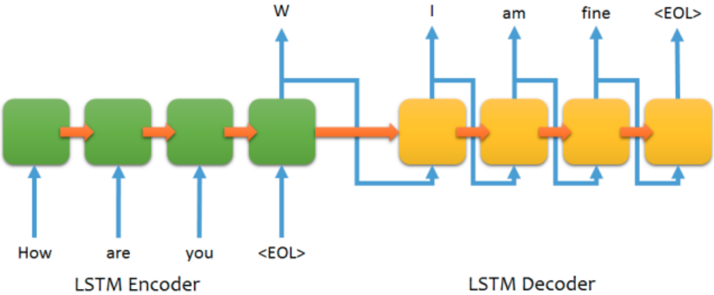
- Simpler and highly explored
- Statistical, Feature-based, Machine Learning, Graph-based

Abstractive TS

The generated summary is a new cohesive text not necessarily present in the original source.

- Hard and challanging
- TS as a neural MT problem; encoder-decoder paradigm

Abstractive TS Problems



Problems

- Generated summary not always meaningful
- Ambiguity to distinguish rare and unknown words
- Grammar errors in the generated summaries

Ideas to Explore

- Infuse prior knowledge like POS tagging
- Infuse other hand-crafted linguistic features
- Inject relational semantic knowledge
- Explore other networks structures or architectures
- ...?

Questions or Suggestions...?

Thank You...☺