Learning Word Vectors for Sentiment Analysis (2011 / 294)

Abstract:

Unsupervised vector-based approaches to semantics can model rich lexical meanings, but they largely fail to capature sentiment information that is central to many word meanings and important for a wide range of NLP tasks. We present a model that uses a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information as well as rich sentiment content. The proposed model can leverage both continuous and multi-dimensional sentiment information as well as non-sentiment annotations. We instantiate the model to unitlize the document-level sentiment polarity annotations present in many online documents. We evaluate the model using small, widely used sentiment and subjectivity copora and find it out-performs several previously introduced methods for sentiment classification. We also introduce a large dataset of movie reviews to serve as a more robust benchmark for work in this area.

Dataset:

- Pang and Lee Movei Review Dataset (2004)
- IMDB (Internet Movie Database) with balanced positive and negative classes

Method:

- Word vectors without traditional stopword removal to preserve sentiment components. Stemming not applied, and non-word tokens kept.
- Model word probabilities conditioned on topic mixture variable.
- MLE (maximum likelihood estimate) for unlabeled documents and MAP (maximum a priori) for topic mixture variable. (unsupervised)
- Logistic regression for sentiment classification (supervised)
- 10-fold cross-validation
- Final objective funciton

$$\nu||R||_F^2 + \sum_{k=1}^{|D|} \lambda||\hat{\theta}_k||_2^2 + \sum_{i=1}^{N_k} \log p(w_i|\hat{\theta}_k; R, b)$$

$$+ \sum_{k=1}^{|D|} \frac{1}{|S_k|} \sum_{i=1}^{N_k} \log p(s_k|w_i; R, \psi, b_c). \qquad (11)$$

Evaluation:

- Both models (w/wo sentiment term) perform better than LSA.
- Improvement over the bag-of-word baseline.

Features	PL04	Our Dataset	Subjectivity
Bag of Words (bnc)	85.45	87.80	87.77
Bag of Words (bΔt'c)	85.80	88.23	85.65
LDA	66.70	67.42	66.65
LSA	84.55	83.96	82.82
Our Semantic Only	87.10	87.30	86.65
Our Full	84.65	87.44	86.19
Our Full, Additional Unlabeled	87.05	87.99	87.22
Our Semantic + Bag of Words (bnc)	88.30	88.28	88.58
Our Full + Bag of Words (bnc)	87.85	88.33	88.45
Our Full, Add'l Unlabeled + Bag of Words (bnc)	88.90	88.89	88.13
Bag of Words SVM (Pang and Lee, 2004)	87.15	N/A	90.00
Contextual Valence Shifters (Kennedy and Inkpen, 2006)	86.20	N/A	N/A
tf.Δidf Weighting (Martineau and Finin, 2009)	88.10	N/A	N/A
Appraisal Taxonomy (Whitelaw et al., 2005)	90.20	N/A	N/A

Table 2: Classification accuracy on three tasks. From left to right the datasets are: A collection of 2,000 movie reviews often used as a benchmark of sentiment classification (Pang and Lee, 2004), 50,000 reviews we gathered from IMDB, and the sentence subjectivity dataset also released by (Pang and Lee, 2004). All tasks are balanced two-class problems.

(REF) Emotions from text machine learnig for text-based emotion prediction (2005/78)

Abstract:

In addtion to information text contains attitudinal, and more specifically, emotional content. This paper explores the text-based emotion prediction problem empirically, using supervised machine learnig with the SNoW learning architecture. The goal is to classify the emotional affinity of sentences in the narrative domain of children's fairy tales, for subsequent usage in appropriate expressive rendering of text-to-speech synthesis. In initial experiments on a preliminary data set of 22 fairy tales show encouraging results over a naive baseline and BOW approach for classifcation of emotional versus non-emotional contents, with some dependency on parameter tuning. We also discuss rsults for a tripartie model which covers emotional valence, as well as feature set alternations. In addition, we present plans for a more cognitively sound sequential model, taking into consideration a larger set of basic emotions.

NO WORD VECTOR

(REF) Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses (2006/285)

Abstract:

Many of the tasks required for semantic tagging of phrases and texts rely on a list of words annotated with some semaantic features. We present a method for extractin sentiment-bearing adjectives from WordNet using the Sentiment Tag Extraction Program (STEP). We did 58 STEP runs on unique non-intersecting seed lists drawn from manually annotated list of positive and negative adjectives and evaluated the results against other manually annotated lists. The 58 runs were then collapsed into a single set o f7813 unique words. For each word we computed a Net Overlap Score by subtracting the total number of runs assigning this word a negative sentiment from the total of the runs that consider it positive. We demonstrate that Net Overlap Score can be used as a measure of the words degree of membership in the fuzzy category of sentiment: the core adjectives, which had the highest Net Overlap Scores, were identified most accurately both by STEP and by human annotators, while the words on the periphery of the category had the lowest scores and were associated with low rates of inter-annotator agreement.

NO WORD VECTOR

(REF) A neural probabilistic language model (2003/145)

Abstract:

A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language. This is intrinsically difficult because of the curse of dimensionality: a word sequence on which the model will be tested is likely to be different from all the word sequences seen during training. Traditional but very successful approaches based on n-grams obtain generalization by concatenating very short overlapping seqquences seen in the training set. We propose to fighting the curse of dimensionality by **learning a distributed representation** for words which allows each training sentence to inform the model about an exponential number of semantically neighboring sentences. The model learns simultaneously (1) a distibuted representation for each word along with (2) the probability function for word sequences, expressed in terms of these representations. Generalization is obtained because a sequence of words that has never been seen before gets high probability if it is made of words that are similar (in the sense of having a nearby representation) to words forming an already seen sentence. Training such large models (with millions of parameters) within a reasonable time is iteself a significant challenge. We report on experiments using **neural networks for the probability function**, showing on two text corpora that the proposed approach significantly **improves on state-of-the-art n-gram models**, and that the proposed approach allows to take advantage of longer contexts.

Dataset:

- the Brown corpus
- Associated Press (AP) News from 1995 and 1996.

TO BE CONTINUED...

(REF) Latent dirichlet allocation (2003/2933)

Abstract:

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data

such as text corpora. LDA is a three-level **hierarchical Bayesian model**, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. We present efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

Dataset:

the TREC AP corpus (Harman, 1992)

Model:

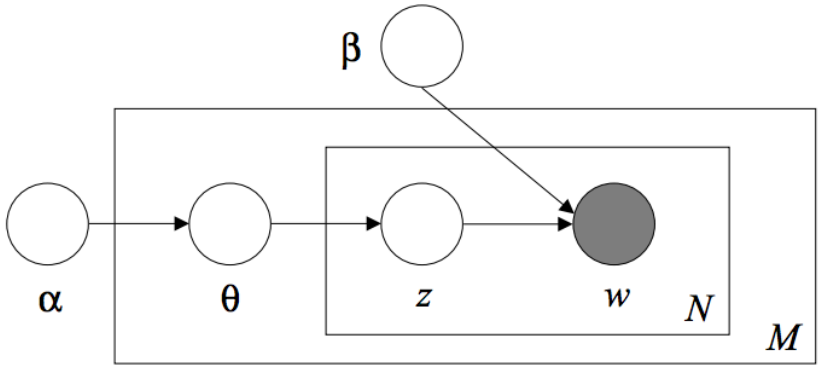

Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

TO BE CONTINUED...

(REF) A unified architecture for natural language processing: deep neural networks with multitask learning (2008/124)

Abstract:

We describe a single convolutional neural network architecture that, given a sentence, outputs a host of language processing predictions: part-of-speech tages, chunks, named entity tags, semantic roles, semantically similar words and the likelihood that the sentence makes sense (grammatically and semantically) using a language model. The entire network is trained jointly on all these tasks using weight-sharing, an instance of multitask learning. All the tasks use labeled data except the language model which is learnt from unlabeled text and represents a novel form of semi-supervised learning for the shared tasks. We show how both multitask learning and semi-supervised learning improve the generalization of the shared tasks, resulting in state-of-the-art performance.

Dataset:

Sections 02-21 of the PropBank dataset version 1 (about 1 million words) for training and Section 23 for testing as standard in all SRL experiments. POS and chunking tasks use the same data split via the Penn TreeBank. NER labeled data was obtained by running the Stanford Named Entity Recognizer on Wikipedia.

Method:

- Part-Of-Speech Tagging; Chunking, labeling sentence segments or phrases; Named Entity Recognition; Semantic role labeling;
- Word vectors with neural network for multitasking learning

Evaluation:

Table 2. A Deep Architecture for SRL improves by learning auxiliary tasks that share the first layer that represents words as wsz-dimensional vectors. We give word error rates for wsz=15, 50 and 100 and various shared tasks.

	wsz=15	wsz=50	wsz=100
SRL	16.54	17.33	18.40
SRL + POS	15.99	16.57	16.53
SRL + Chunking	16.42	16.39	16.48
SRL + NER	16.67	17.29	17.21
SRL + Synonyms	15.46	15.17	15.17
SRL + Language model	14.42	14.30	14.46
SRL + POS + Chunking	16.46	15.95	16.41
SRL + POS + NER	16.45	16.89	16.29
SRL + POS + Chunking + NER	16.33	16.36	16.27
SRL + POS + Chunking + NER + Synonyms	15.71	14.76	15.48
SRL + POS + Chunking + NER + Language model	14.63	14.44	14.50

(REF) Joint parsing and named entity recognition (2009/32)

Abstract:

For many language technology applications, such as question answering, the overall system runs several independent processors over the data (such as a named entity recognizer, a coreference system, and a parser). This easily results in inconsistent annotations, which are harmful to the performance of the aggregate system. We begin to address this problem with **a joint model of aprsing and named entity recognition, based on a discriminative feature-based constituency parser**. Our model produced a consistent output, where the named entity spans do not conflict with the phrasal spans of the parse tree. The joint representation also allows the information from each type of annotation to improve performance on the other, and, in experiments with the OntoNotes corpus, we found improvements of up to 1.36% absolute F1 for parsing, and up to 9.0% F1 for named entity recognition.

Dataset:

LDC2008T04 OntoNotes Release 2.0 corpus

JUST CHUNKING AND TAGGING

(REF) Seeing stars when there aren't many stars: graph-based semi-

supervised learning for sentiment categorization (2006/56)

Abstract:

We present a graph-based semi-supervised learning algorithm to address the sentiment analysis task of rating inference. Given a set of documents (e.g. movie reviews) and accompanying ratings, the task calls for inferring numerical ratings for unlabeled documents based on the perceived sentiment expressed by their text. In particular, we are interested in the **situation where labeled data is scarce**. We place this task in the semi-supervised setting and demonstrate that considering unlabeled reviews in the learning process can improve rating-inference performance. We do so by **creating a graph on both labeled and unlabeled data to encode certain assumptions** for this task. We then solve an optimization problem to obtain a smooth rating function over the whole graph. When only limited labeled data is available, this method achieves significantly better predictive accuracy over other methods that ignore the unlabeled examples during training.

Dataset:

http://www.cs.cornell.edu/people/pabo/movie-review-data/ used in (Pang and Lee, 2005)

Method:

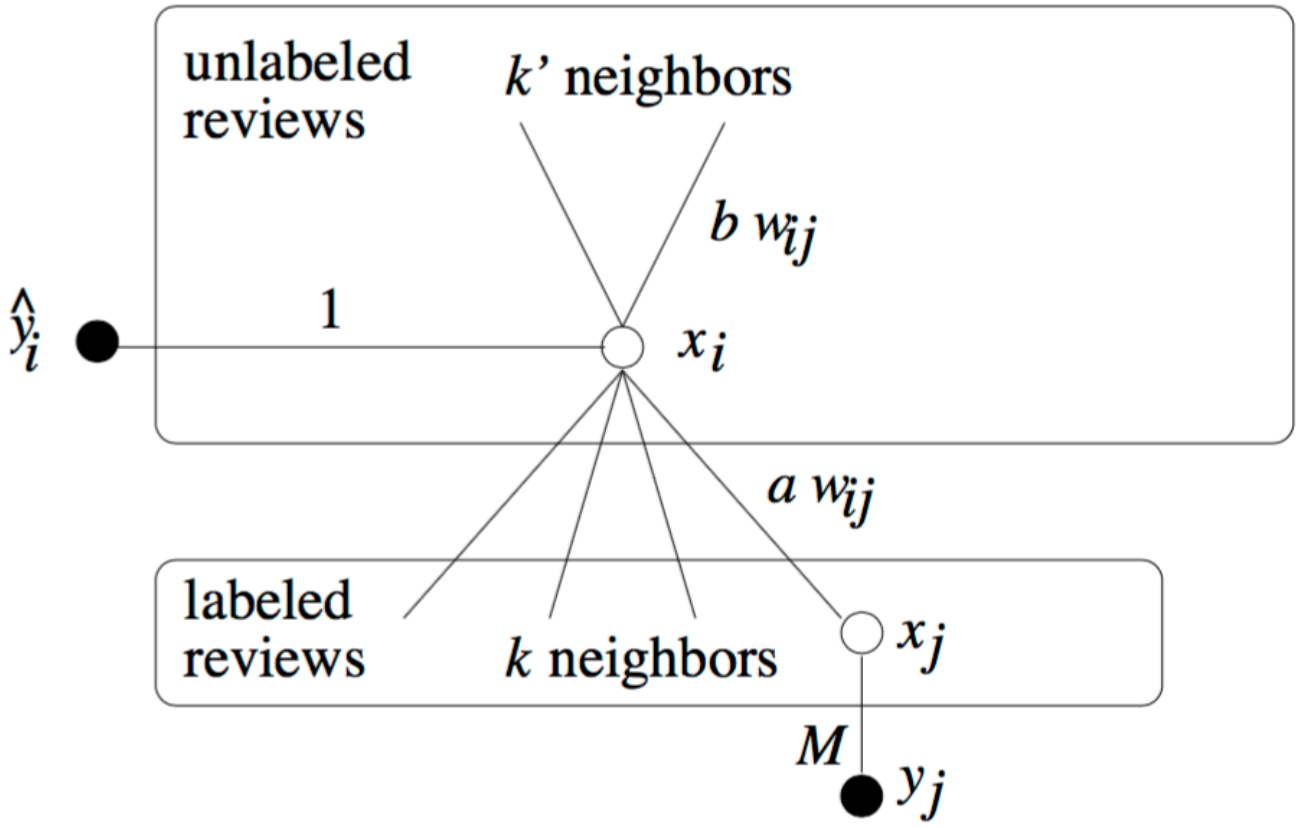

Figure 1: The graph for semi-supervised rating inference.

Evaluation:

Achieved better performance than all other methods in all four author corpora

(REF) Joint sentiment/topic model for sentiment analysis (2009/97)

Abstract:

Sentiment analysis or opinion mining aims to use automated tools to detect subjective information such as opinions, attitudes, and feelings expressed in text. This paper proposes a novel probabilistic modeling framework based on LDA, called joint sentiment/topic model (JST), which detects sentiment and topic simultaneously from text. Unlike other machine learning approaches to sentiment classification which often require labeled corpora for classifier training, the proposed JST model is fully unsupervised. The model has been evaluated on the movie review dataset to classify the review sentiment polarity and minimum prior information have also been explored to further improve the sentiment classification accuracy. Preliminary experiments have shown promising results achieved by JST.

Dataset:

dataset consists of two cat- egories of free format movie review texts, with their overall sentiment polarity labeled either positive or negative.

Method:

a joint sentiment/topic (JST) model by adding an additional sentiment layer between the docu- ment and the topic layer.

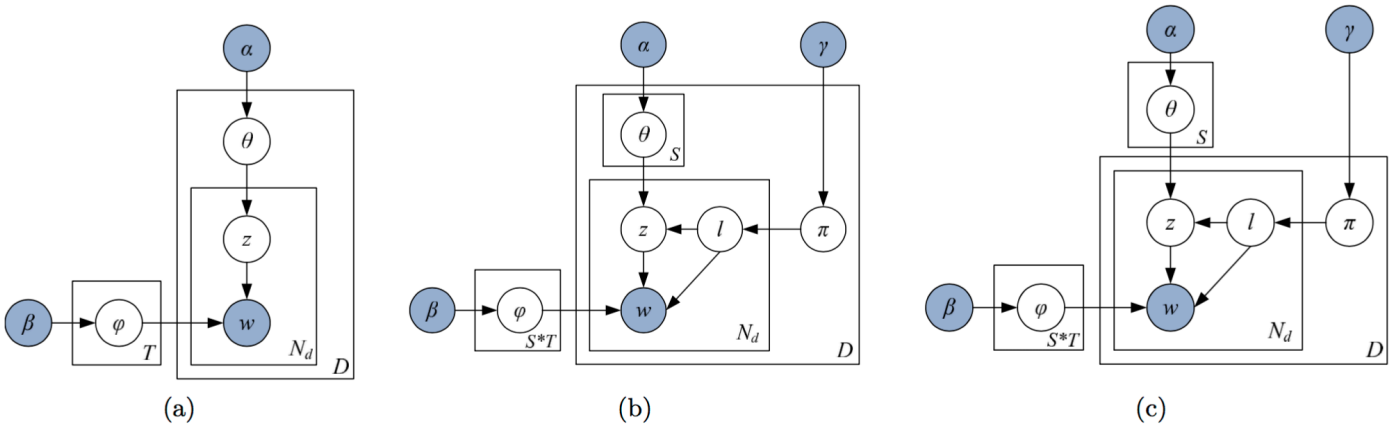

Figure 1: (a) LDA model; (b) JST model; (c) Tying-JST model.

(REF) Thumbs Up?: sentiment classification using machine learning techniques (2002/890)

Abstract:

We consider the problem of classifying documents not by topic, but by overall sentiment, e.g., determing whether a review is positive or negative. Using **movie reviews as data**, we find that standard machine learning techniques definitively outperform human-produced baselines. However, the three machine learning methods we employed (**Naive Bayes, maximum entropy classification, and support vector machines) do not perform as well on sentiment classification as on traditional topic-based categorization**. We conclude by examining factors that make the sentiment classification problem more challenging.

TO BE CONTINUED...

(REF) Word representations: a simple and general method for semi-supervised learning (2010/87)

Abstract:

If we take an existing supervised NLP system, a simple and gernal way to improve accuracy is to use unsupervised word representations as extra word features. We evaluate Brown clusters, Collobert and Weston (2008) embeddings, and HLBL (Mnih & Hinton, 2009) embeddings of words on both NER and chunking. We use near state-of-the-art supervised baselines, and find that each of the three word representations improves the accuracy of these baselines. We find further improvements by combinng different word representations.

TO BE CONTINUED...

Baseline and Bigrams: Simple, Good Sentiment and Topic Classification (2012/17)

Abstract:

Variants fo Naive Bayes (NB) and Support Vector Machines (SVM) are often used as baseline methods for text classification, but their performance varies greatly depending on the model variant, features used and task/dataset. We show that: (i) the **inclusion of word bigram features gives consistent gains on sentiment analysis tasks**; (ii) for short snippet sentiment tasks, NB actually does better than SVMs (while for longer documents the opposite result holds); (iii) a simple but novel **SVM variant using NB log-count ratios as feature values** consistently performs well across tasks and datasets. Based on these observations, we identify simple NB and SVM variants which outperform most published results on sentiment analysis dataesets, sometimes providing a new state-of-the-art performance level.

TO BE CONTINUED...

End-to-End text recognition with convolutional neural networds (2012/121)

Abstract:

Full end-to-end text recognition in nautural images is a challenging problem that has received much attention recently. Traditional systems in this area have relied on elaborate models incorporating carefully hand-engineered features or large amounts of prior knowledge. In this paper, we take a different route and combine the representational power of large, multilayer neural networks together with recent developments in unsupervised feature learning, which allows us to use a common framework to train highly-accurate text detector and character recognizer modules. Then, using only simple off-the-shelf methods, we integrate these two modules into a full end-to-end, lexicon-driven, scene text recognition system that achieves state-of-the-art performance on standard benchmarks, namely Street View Text and ICDAR 2003.

Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification (2014/64)

Abstract:

We present a method that learns word embedding for Twitter sentiment classification in this paper. Most existing algorithms for learning continuous word representations typically only model the syntactic context of words but ignore the sentiment of text. This is problematic for sentiment analysis as they usually map words with similar syntactic context but opposite sentiment polarity such as good and bad, to neighboring word vectors. We address this issue by learning **sentiment specific word embedding (SSWE)**, which encodes sentiment information in the continuous representation of words. Specifically, we develop three **neural networks** to effectively incorporate the supervision from sentiment polarity of text in their loss functions. To obtain large scale training corpora, we learn the sentiment-specific word embedding from mamssive distant-supervised tweets collected by postive and negative emotions. Experiments onn applying SSWE to a benchmark Twitter sentiment classification dataset in SemEval 2013 show that (1) the SSWE feature preforms comparably with hand-crafted freatrures in the top-performed system; (2) the performance is futhre improved by concatenating SSWE with existing feature set.

Dataset:

latest Twitter sentiment clas- sification benchmark dataset in SemEval 2013 (Nakov et al., 2013)

*Evaluation of Word Vector Representations by Subspace Alignment (2015/3)

Abstract:

Unsupervisedly learned word vectors have proven to provide exceptionally effective features in many NLP tasks. Most common intrinsic evaluations of vector quality measure correlation with similarity judgements. However, thesee often correlate poorly with how well the learned represetnations perform as features in downstream evaluation tasks. We present QVEC--a computationally inexpensive intrinsic evaluation measure of the quality of word embeddings based on alignment to a matrix of features extracted from manually crafted lexical resources--that obrains strong correlation with performance of the vectors in a battery of downstream semantic evaluation tasks.

Dataset:

- an existing semantic resource: SemCor(Miller et al. 1993)
- WordNet (Fellbaum 1998)
- WS-353 dataset (Finkelstein et al., 2001), MEN dataset (Bruni et al., 2012), SimLex-999 (Hill et al., 2014) for word similarity test.
- 20 Newsgroups (20NG) dataset for text classification.

Method:

- Construct Word Vectors with annotation dismensions

WORD	NN.ANIMAL	NN.FOOD	...	VB.MOTION
fish	0.68	0.16	...	0.00
duck	0.31	0.00	...	0.69
chicken	0.33	0.67	...	0.00

Table 1: Oracle linguistic word vectors, constructed from a linguistic resource containing semantic annotations.

- Word Vector Evaluation Model

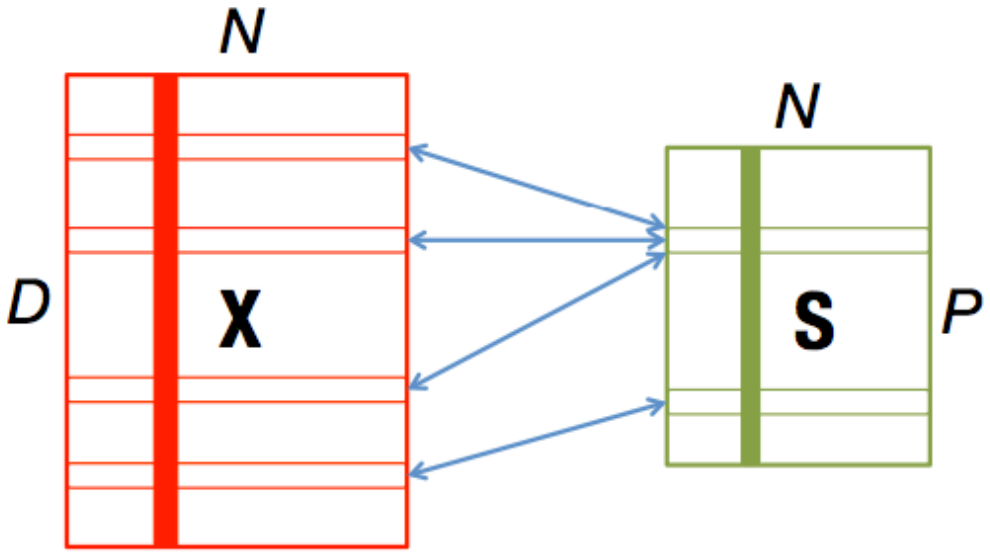

Figure 1: The filled vertical vectors represent the word vector in the word vector matrix **X** and the linguistic property matrix **S**. The horizontal hollow vectors represent the "distributional dimension vector" in **X** and "linguistic dimension vector" in **S**. The arrows show mapping between distributional and linguistic vector dimensions.

Evaluation:

Model	QVEC	Senti
CBOW	40.3	90.0
SG	35.9	80.5
CWindow	28.1	76.2
SSG	40.5	81.2
Attention	40.8	80.1
GloVe	34.4	79.4
GloVe+WN	42.1	79.6
GloVe+PPDB	39.2	79.7
LSA	19.7	76.9
LSA+WN	29.4	77.5
LSA+PPDB	28.4	77.3
Correlation (r)	**0.87**	

Table 2: Intrinsic (QVEC) and extrinsic scores of the 300-dimensional vectors trained using different word vector models and evaluated on the Senti task. Pearson's correlation between the intrinsic and extrinsic scores is $r = 0.87$.

TO BE CONTINUED...

*Generating Overview Summaries of Ongoing Email Thread Discussions (2004/65)

Abstract:

The tedious task of responding to a blacklog of email is one which is familiar to many researchers. As a subset of email management, we address the problem of constructing a summary of email discussios. Specifically, we

examine ongoing discussions which will untilmately culminate in a consensus in a decsion-making process. Our summary provides a snapshot of the current state-of-affairs of the discussio and facilitates a speedy response fromthe user, who might be the bottleneck in some matter being resolved. We present a method which uses the sructure of the thread dialogue and word vector techniques to determine which sentence in the thread shoud be extracted as the main issue. Our solution successfully identifies the sentence containing the issue of the thread being discussed, potentially more informative that subject line.

Dataset:

Archives of Columbia Unversity ACM Student Chapter Committee

Method:

- Combination of traditional vector space techniques and Singular Value Decomposition (SVD).

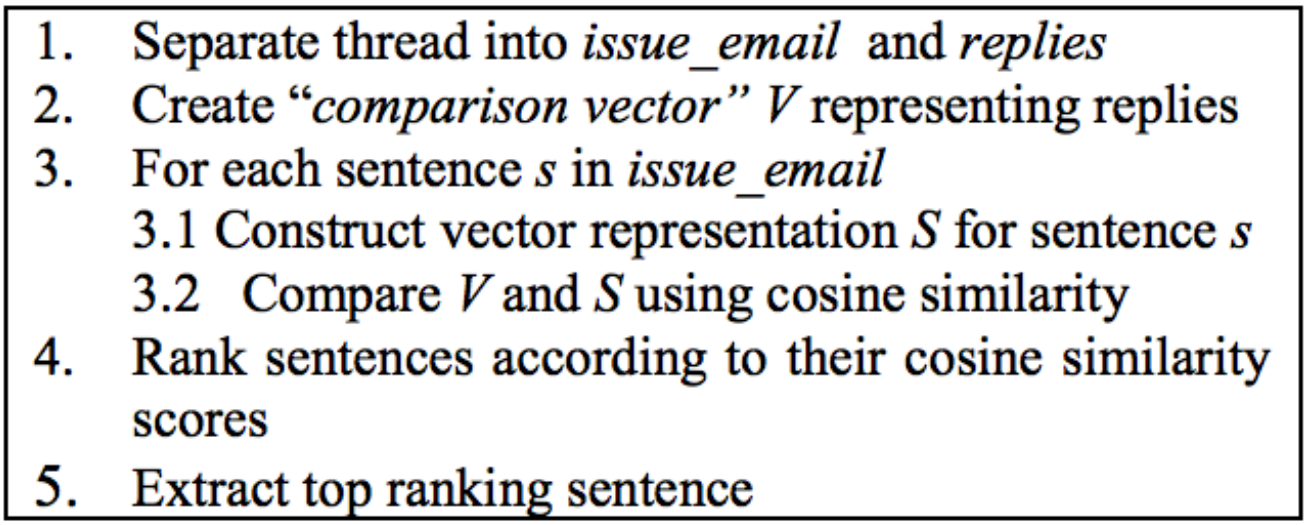

Figure 3. Framework for extracting discussion Issues.

Evaluation:

a combination of simple word vector approaches with singular value decomposition approaches do well at extracting discussion issues.

TO BE CONTINUED...

*Distributed Representations of Words and Phrases and their Compositionality (2013/1133)

Abstract:

The recently introduced continuous Skip-gram model is an efficient method for learning hgih-quality distributed vector representations that capture a large number of precise syntatic and semantic word relationships. In this

paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word represetntations. We also describe a simple alternative to the hierarchical softmax called negative sampling. An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of Canada and Air cannot be easily combined to obtain Air Canada. Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector represenations for millions of phrases is possible.

Dataset:

Google New articles

Method:

- Skip-gram model (Advantage over previous neural network: not involve dense matrix multiplication).
- Subsampling to counter the imbalance between the rare and frequent words.
- Hierarchical softmax.

Evaluation:

- Large amount of training data is crucial to increase the accuracy.
- A big Skip-gram model outperform all previously published word representation methods.

TO BE CONTINUED...

*Efficient Estimation of Word Representations in Vector Space (2013/1205)

Abstract:

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techninques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-art performance on our test set for measuring syntactic and semantic word similarities.

Dataset:

Google New corpus

Models:

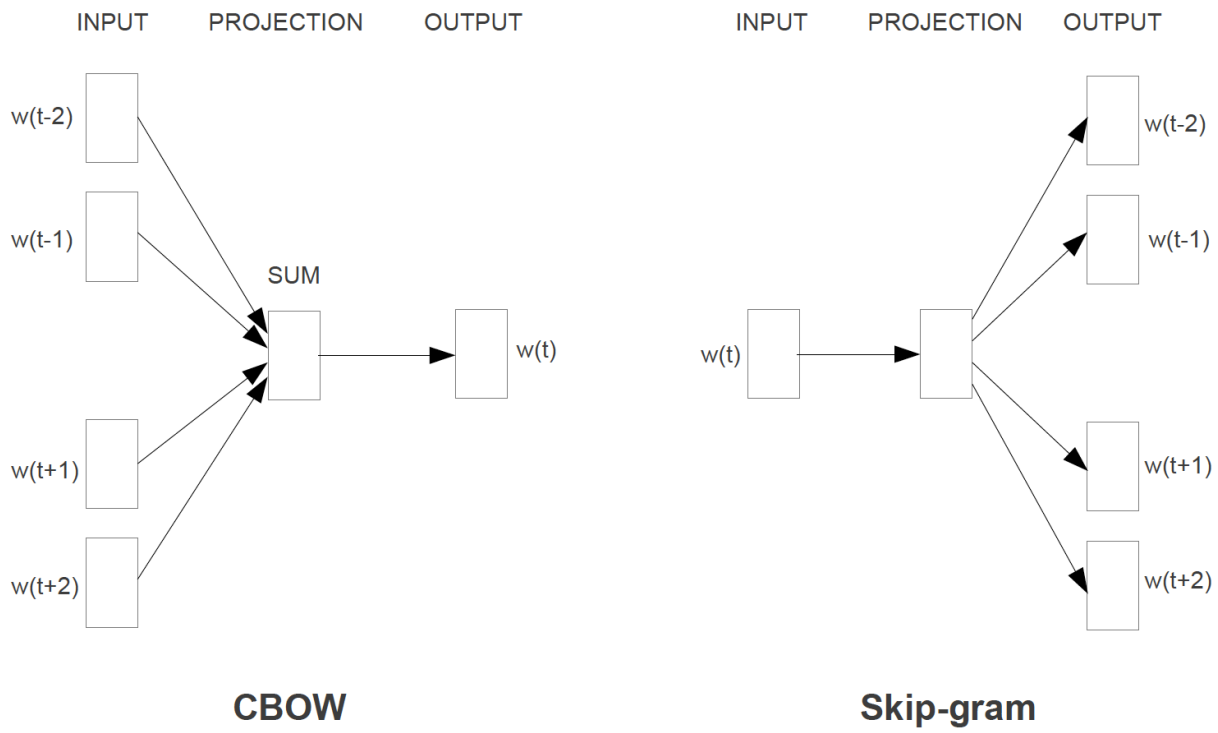

Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

TO BE CONTINUED..