Reading Between the Lines

A Vector Space Model of Language Using Semantic Role Structures

Charley Wu[1] (charleymswu@gmail.com), Marcin Skowron[2], and Paolo Petta[2]
[1]MEi: CogSci, Universität Wien
[2]Austrian Research Institute for Artificial Intelligence (OFAI)

Introduction

In this poster, we present several methods for creating a compositional Vector Space Model (VSM) that can capture the semantic difference between texts based on opposing ideological positions. We compare the effectiveness of these methods by training a Support Vector Machine (SVM) classifier on the vector outputs produced by the model. Traditionally, VSMs have been trained using Latent Semantic Analysis (LSA), which uses the context of the occurence words to learn a representation about their meaning. This is based on the notion that words that occur in the same contexts have a similar meaning. We add to this framework by joining the LSA approach with Semantic Role Labeling (SRL) to provide structured data for the creation of a compositional VSM. The goal is to represent the semantic information of an entire document.

Process

Part 1. Word2Vec - from words to vectors

This is an LSA model we trained on the entire English Wikipedia. Word2Vec allows us to query any word it has learned and return a 300 dimensional word vector. Word vectors are composed in such a way that Cosine Distance is a measure of semantic similarity.
(Complexity: 48hrs, 60 GB)

Part 2. Semantic Role Labeling - adding structure

Using a Deep Neural Network trained on the Wall Street Journal corpus (SENNA), we interpret the relationships between semantic arguments in a text. This provides the structure about how individual words are related to each other.
(Complexity: 24 hrs, 2 GB)

Part 3. TF-IFD Model - which words are important

Term Frequency–Inverse Document Frequency is a measure of how important a given word is to a document in a given corpus. The model was trained on a subset of the CORPS corpus and is used to scale the magnitude of the word vectors to reflect the amount of information it contributes to the document.
(Complexity: 1 min, 3 MB)

Part 4. Compositionality - putting it together

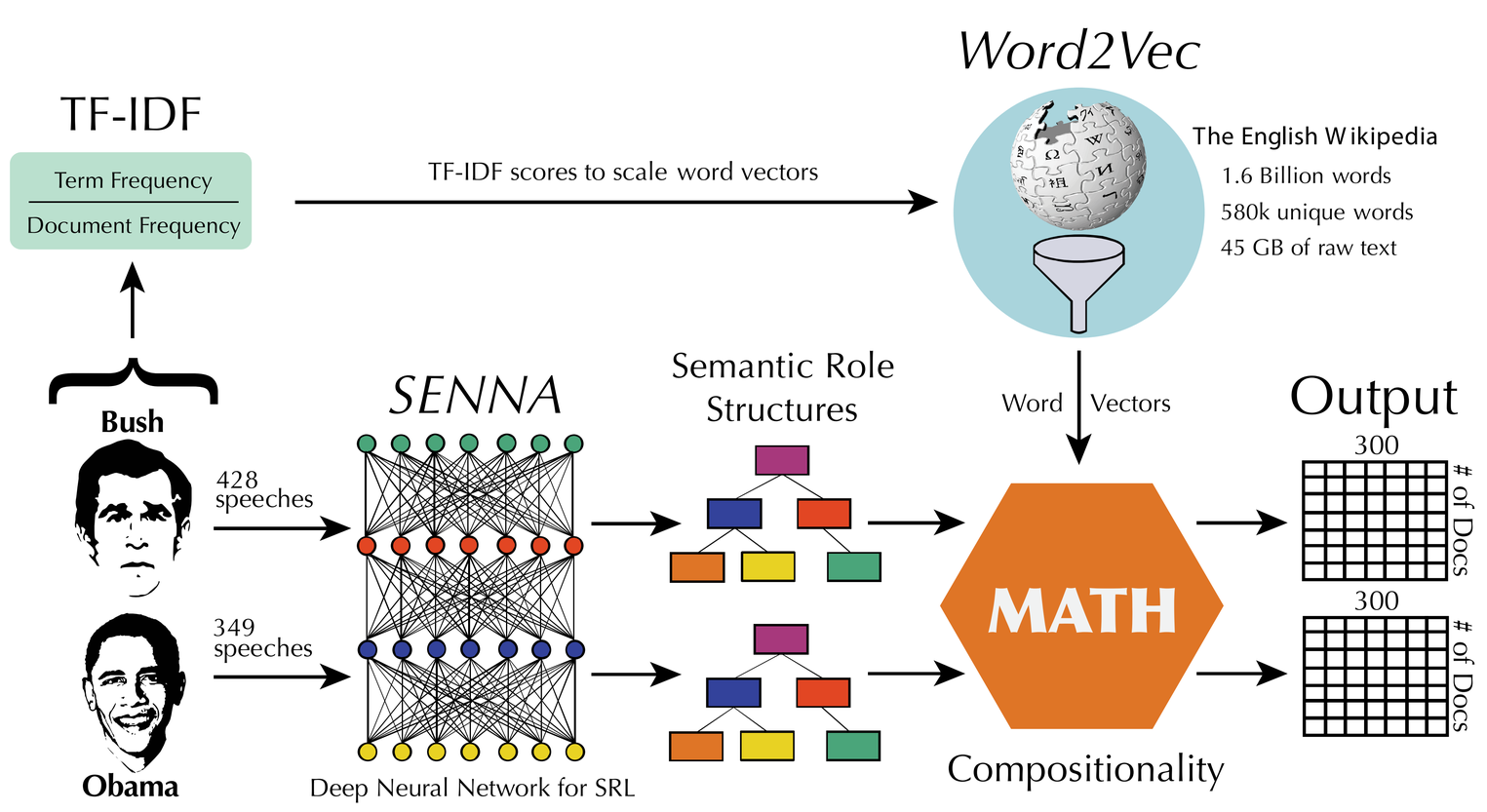
For each semantic role structure in a document, we translated each word token into a vector using Word2Vec. After scaling each word vector by it's TF-IDF score, we performed one of 6 different compositionality methods to produce a single vectorized representation of a document. Each corpus is a single matrix where each row is a document vector. (Complexity: 2 hrs, 2 GB)

Part 5. Classification - evaluating the methods

The two corpus matrices were joined and randomly shuffled.
To visualize the data in 2D, we projected the matrix against Fisher's Linear Discriminant. To compare the different compositionality methods, we trained a SVM classifier with a Gaussian Kernel for each. The performance was measured by performing 5-fold cross validation and by plotting an ROC curve.
(Complexity: 10 mins, 2 GB)

CORPS: Political Speeches

The CORpus of tagged Political Speeches (CORPS) is a large collection of annotated speeches from various political and public figures. Originally compiled for research in persuasion and audience reaction, we have discarded the added annotation and trained our language model from the basic text alone. Choosing a collection of 777 speeches from the previous and current Presidents of the US, we compared the effectiveness of different compositionality methods based on the seperability of the two classes of documents.



Technical Glossary

Compositionality: The use of word order or syntax to construct a vectorized representation of a phrase
Fisher Projection: a Dimensionality Reduction method
LSA: Latent Semantic Analysis
ROC: Reciever Operating Characteristic; used to evaluate the performance of a binary classifier
SRL: Semantic Role Labeling; detecting the semantic arguments associated with a predicate/verb in a sentence
SVM: Support Vector Machine; a supervised learning algorithm that can perform non-linear classification
TF-IDF: Term Frequency - Inverse Document Frequency
VSM: Vector Space Model; mathematical method of representing text documents

Compositionality Methods

1. Additive

$$\sum_{i=1}^{N} \vec{w}_{ij}$$

2. Multiplicative

$$\prod_{i=1}^{N} \vec{w}_{ij}$$

3. Circular Convolution

$$\sum_{i=1}^{N} \vec{w}_{i-1} \circledast \vec{w}_i$$

4. Circular Convolution /Additive

$$\vec{w}_v \circledast \sum_{i=1}^{N_{a0}} \vec{w}_{ij}^{a0} + \sum_{i=1}^{N_{A1}} \vec{w}_{ij}^{a1} \circledast \vec{w}_v + \sum_{i=1}^{N_{rem}} \vec{w}_{IJ}$$

5. Tensor Product/Additive

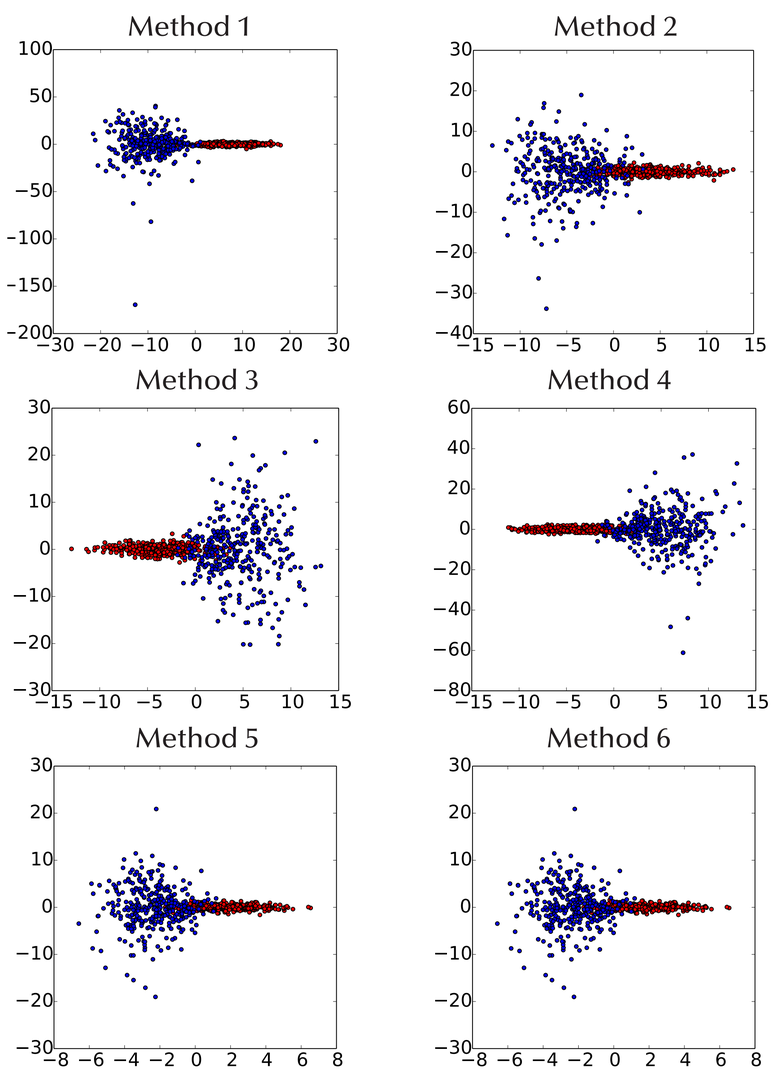
$$\sum_{i=1}^{N} \left[\vec{w}_v \odot \sum_{i=1}^{N_{a0}} \vec{w}_{ij}^{a0} + \sum_{i=1}^{N_{a1}} \vec{w}_{ij}^{a1} \odot \vec{w}_v \right]_{ij} + \sum_{i=1}^{N_{rem}} \vec{w}_{ij}$$

6. Circular Convolution/Dot Product

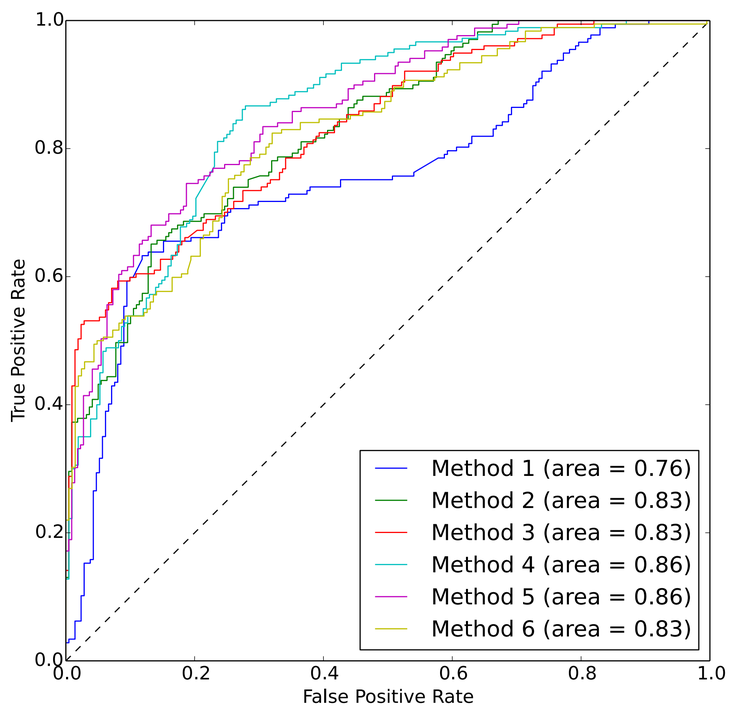
$$\left[\vec{w}_v \circledast \sum_{i=1}^{N_{a0}} \vec{w}_{ij}^{a0} \right] \cdot \left[\sum_{i=1}^{N_{a1}} \vec{w}_{ij}^{a1} \circledast \vec{w}_v \right] + \sum_{i=1}^{N_{rem}} \vec{w}_{ij}$$

Fisher Projections

B.H. Obama
G.W. Bush



ROC Curve



5-Fold Cross Validation

Method 1: SVM classifier scores **0.62 (+/- 0.04)**
Method 2: SVM classifier scores **0.79 (+/- 0.03)**
Method 3: SVM classifier scores **0.76 (+/- 0.05)**
Method 4: SVM classifier scores **0.84 (+/- 0.06)**
Method 5: SVM classifier scores **0.78 (+/- 0.02)**
Method 6: SVM classifier scores **0.79 (+/- 0.05)**

Discussion

We set Method 1 as a *baseline* becauses it is essentially a non-compositional VSM. It makes almost no use of the semantic structure provided by SENNA and is equivalent to a vanilla Word2Vec approach with the addition of TF-IDF vector scaling. Methods 2 and 3 also make little use of SRL structure but have different mathematical approaches.

Methods 4-6 specifically use the relationship of the {verb} with {A0} ie. agent and {A1} ie. patient. Based on the discrimination task, Method 4 produced the best performance, while Methods 5-6 were not significantly different from the non-compositional Methods 2 and 3. However, each of the Methods 2-6 performed better than the baseline.

The next step in this direction of research will be to train a *Generative Classifier* such as a Gaussian Mixture Model or Latent Direchlet Allocation. This will allow for the generation of new text, using probabalistic sampling to recompose a semantic vector into individual words in a semantic role structure.

The generation of new text is where the more structured compositional Methods (ie. 4-6) are expected to prove useful, because they narrow the scope of possible word combinations that can be reconstructed from a given semantic vector. For Methods 4-6, meeting the baseline in this discrimination task indicate that they are viable options as compositional representations of text.

Acknowledgements

CORPS was provided for research purposes by Marco Guerini
OFAI is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology.
SENNA was written by Ronan Collobert and is availble through non-commercial licence by NEC Labs America
Word2Vec was implemented through Gensim, which is written by Radim Rehurek