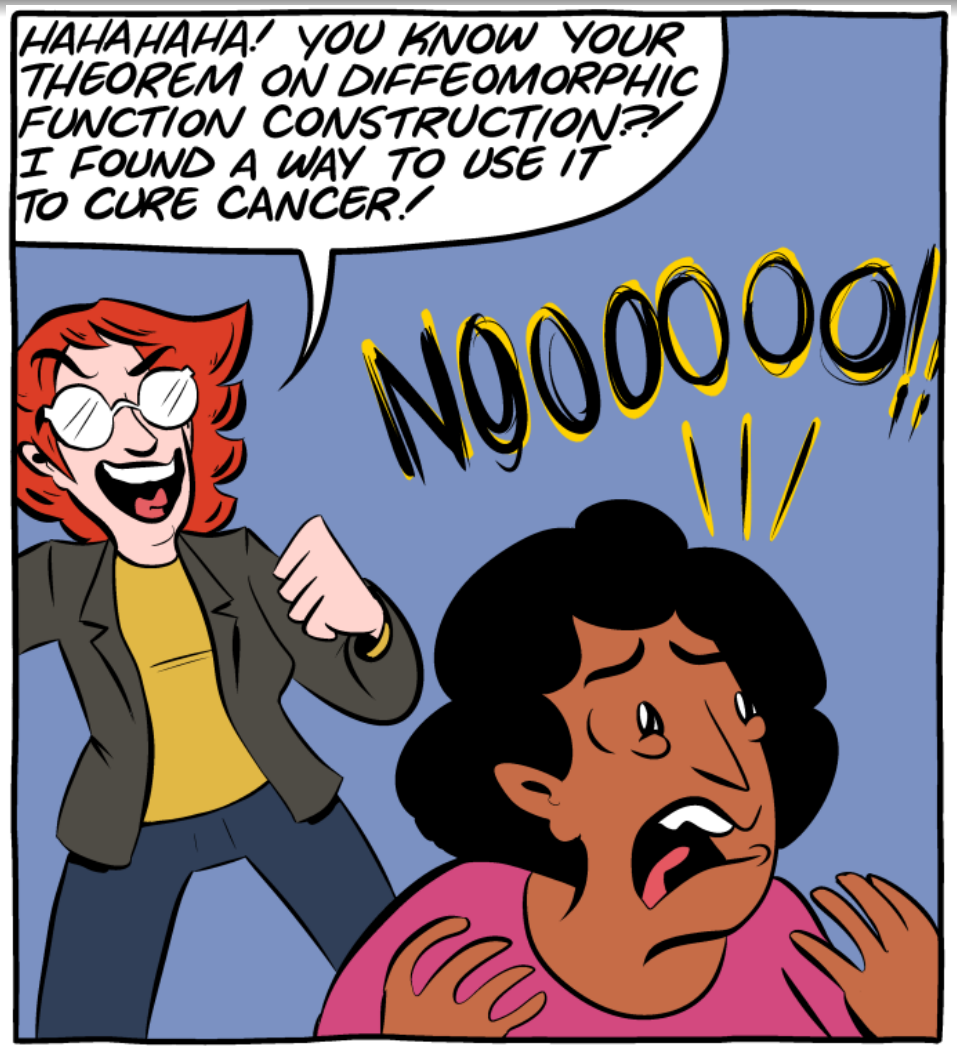
A summary of NLP techniques: Torwards Deep Learning.

Felipe Pérez

June 26, 2017



Funtime Activity:
Forcibly converting pure mathematicians into applied mathematicians.

What you should know/do to get the most out of this:

- Read.
- Try it out.
- Read again.
- Basic Linear Algebra: vector spaces, linear transformation, SVD and its geometric meaning.
- Basic Probability: typical distributions, interpretations, independence, conditional, Bayes' Theorem.
- Basic ideas behind gradient descent, neural networks, and back propagation.
- Python: Numpy, Pandas, Tensorflow, Keras.

NLP IS HARD!

The hard tasks - Translation

- Translation: Google translate used statistical translation until 2016 when they changed to deep learning approach, using LSTM. But there's still work to do!

Example

Ιθάκη - Καβάφης

Σα βγεις στον πηγαιμό για την Ιθάκη,
να εύχεσαι νάναι μακρύς ο δρόμος,
γεμάτος περιπέτειες, γεμάτος γνώσεις.
Τους Λαιστρυγόνας και τους Κύκλωπας,
τον θυμωμένο Ποσειδώνα μη φοβάσαι,
τέτοια στον δρόμο σου ποτέ σου δεν θα βρεις,
αν μέν' η σκέψις σου υψηλή, αν εκλεκτή
συγκίνησις το πνεύμα και το σώμα σου αγγίζει.
Τους Λαιστρυγόνας και τους Κύκλωπας,
τον άγριο Ποσειδώνα δεν θα συναντήσεις,
αν δεν τους κουβανείς μες στην ψυχή σου,
αν η ψυχή σου δεν τους στήνει εμπρός σου.

Example

Ithaca - Cavafy

When you go to Ithaca,
To long wait for the long road,
Full of adventures, full of knowledge.
Laistrogens and Cyclops,
The angry Poseidon is not afraid,
You will never find yourself on your way,
If your thought is high, if you are chosen
Touch your spirit and your body touches.
Laistrogens and Cyclops,
The wild Poseidon will not meet,
If you do not kiss them in your soul,
If your soul does not set them in front of you.

Example

Ithaca - | Cavafy

As you set out for Ithaka
hope the voyage is a long one,
full of adventure, full of discovery.
Laistrygonians and Cyclops,
angry Poseidon—don't be afraid of them:
you'll never find things like that on your way
as long as you keep your thoughts raised high,
as long as a rare excitement
stirs your spirit and your body.
Laistrygonians and Cyclops,
wild Poseidon—you won't encounter them
unless you bring them along inside your soul,
unless your soul sets them up in front of you.

Ithaca - Cavafy

When you go to Ithaca,
To long wait for the long road,
Full of adventures, full of knowledge.
Laistrogens and Cyclops,
The angry Poseidon is not afraid,
You will never find yourself on your way,
If your thought is high, if you are chosen
Touch your spirit and your body touches.
Laistrogens and Cyclops,
The wild Poseidon will not meet,
If you do not kiss them in your soul,
If your soul does not set them in front of you.

The hard tasks - Semantic Analysis

- What is the meaning of [INSERT NL OBJECT HERE]? Currently using: Vector Space Models (VSM). Idea: Make data into vectors and understand their semantic meaning out of the relation.

Example

- What is a "table"?
- What is "one"? One is the equivalence class of the set $\{\emptyset\}$, under the equivalence relation on sets given by bijections.
- What is "time"?

Vector Space Models

Transform the data into vectors! **How?**

- Make each character a vector or,
- Make each word a vector or,
- Make each sentence a vector or,
- Make each document a vector.

One Hot encodding

We transform the different data (characters, words, etc.) into different axis in a large vector space.
That is, if we have 1000 words, then we would obtain the standard basis for a 1000-dimensional vector space.
Problem: This does not encode the relations between different words.

Capturing their relations

This is where the co-occurrence matrix comes into play. This matrix will keep track of the number of times two words occur together. For example, if mouse=$(1, 0, \ldots, 0)$ and cat=$(0, 1, 0, \ldots, 0)$, with the words cat and mouse appear together in 153 sentences, then the co-occurrences matrix C is such that

$$C(1, 2) = C(2, 1) = 153.$$

Assumption

A big part of NLP is based on

Hypothesis

Linguistic items with similar distributions have similar meanings.

Warning:

This is a somewhat strong hypothesis. It creates problem with homonyms. Consider:

"Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo"

This is a correct english sentence meaning:

"Bison from Buffalo, New York who are intimidated by other bison in their community also happen to intimidate other bison in their community."

Capturing their relations

Key idea

Instead of using the one-hot encoding we can use the rows of this matrix!

Problem

Wait! This is a huge matrix!

- In most languages there are between 100,000 and 500,000 words. The matrix, therefore, would be huge.
- Most entries are zero (Sparcity).

An approach

Build the best possible approximation for the matrix C in a lower dimensional vector space. This is achieved in two steps:

1. Find the Singular Value Decomposition (SVD).
2. Use SVD to find a lower dimentional approximation.

SVD

Singular Value Decomposition

Any matrix A can be factored as

$$A = U\Sigma V^T,$$

where U and V are orthogonal matrices, and Σ is a diagonal matrix with non-negative real numbers in the diagonal.

Approximation

Choose the columns of U associated to the k largest eigenvalues, and call this matrix U_k. Then the rule

$$\text{word} \to w = \text{one-hot} \to U_k^T \cdot w$$

That is, U_k gives an embedding of the words to a lower dimentional vector space (of dim $= k$).

This is the best approximation for the cosine distance of vectors.

Problems with this approach

This approach still has several semantic problems as well as practical ones.

- Still high dimentional matrix.
- Quadratic cost on computing SVD.
- Some words appear more often ('The','I',etc.).
- Homonomies.
- Size of sentences may not be large enough to capture meaning.
- Depend on the source of the sentences.

Using probability

Note that in the previous approach we were using how often something happens as a feature. Let's try to switch the way we look at it a bit and now consider the probability of something happening as the feature.

Continuous Bag of Words Model

In this model we want to predict a word out of its context.

$$INPUT = \text{CONTEXT} \rightarrow OUTPUT = \text{WORD}.$$

For example

___ is the best food in the world. $\rightarrow$ Icecream

More precisely

We want to predict word_c, given the m previous words, $\text{word}_{c-m}, \ldots, \text{word}_{c-1}$ and the m subsequent words, $\text{word}_{c+1}, \ldots, \text{word}_{c+m}$. That is:

$$Input = \{\text{word}_{c-m}, \ldots, \text{word}_{c-1}, \text{word}_{c+1}, \ldots, \text{word}_{c+m}\}$$

$$\downarrow$$

$$Output = \text{word}_c$$

What's the arrow?

We break it into steps, let N the number of distinct words:

1. Take each word $\text{word}_{c-m}, \ldots, \text{word}_{c+m}$ and use one-hot encoding to obtain $w_{c-m}, \ldots, w_{c+m}$.
2. Use a VSM to embed into a smaller n-dimentional space, this creates a matrix U of size $N \times n$. We get

$$u_{c-m} = U \cdot w_{c-m}, \ldots, u_{c+m} = U \cdot w_{c+m}$$

What's the arrow?

3. Take the average of the vectors

$$u_c = \frac{u_{c-m} + \ldots + u_{c+m}}{2m}.$$

4. Go back to the one-hot space, this creates a matrix V of size $n \times N$. We get $v_c = V \cdot u_c$.

What's the arrow?

5. Interpret the vector v_c as probabilities. This is done via a softmax function. $\hat{y}_c = \text{softmax}(v_c)$.

In summary:

$$\text{Input} \xrightarrow{\text{one-hot}} \{w_i\} \xrightarrow{U} \{u_i\} \xrightarrow{\text{average}} u_c \xrightarrow{V} v_c \xrightarrow{\text{softmax}} \hat{y}_c \longrightarrow \text{Output}$$

Finding the right embedding

Cool, but...

... How do we find U and V?

We haven't used word_c!

Finding the right embedding

Let y_c be the one-hot encoding of word$_c$. We want to find U and V such that the difference on the distribution obtained by $\hat{y}_c$ and the real distribution y_c are minimized. We use cross-entropy for this. That is, we want to minimize

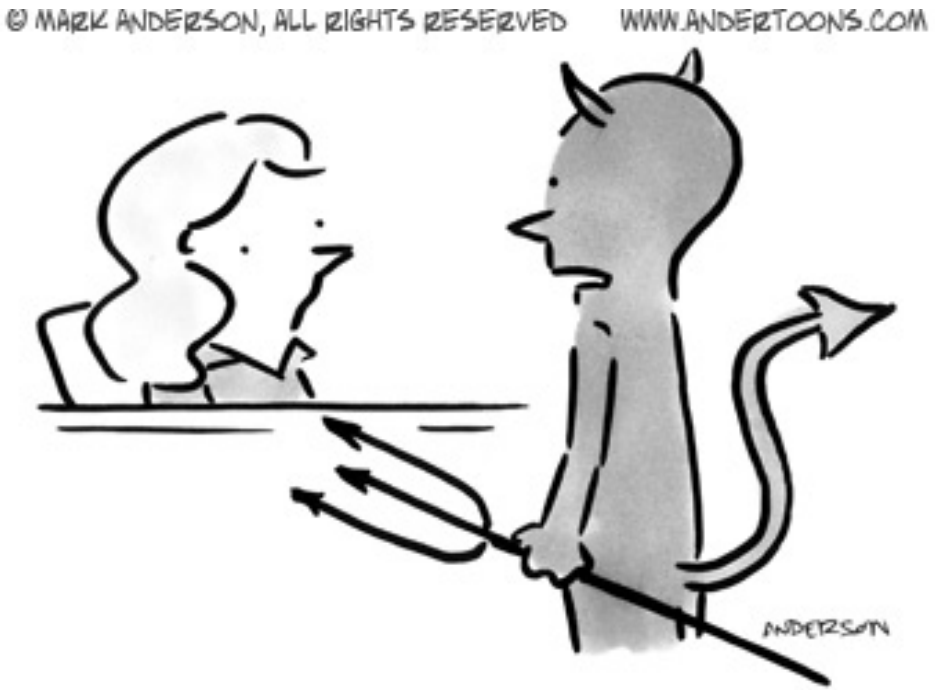
$$H(\hat{y}, y) = -\sum_c \sum_i (y_c)_i \log((\hat{y}_c)_i)$$

And this can be done via gradient descent!

Soon

- Glove.
- RNN, LSTM,GRU
- Deeper models

Questions



"I'm here about the details."

Social time next

Thanks!