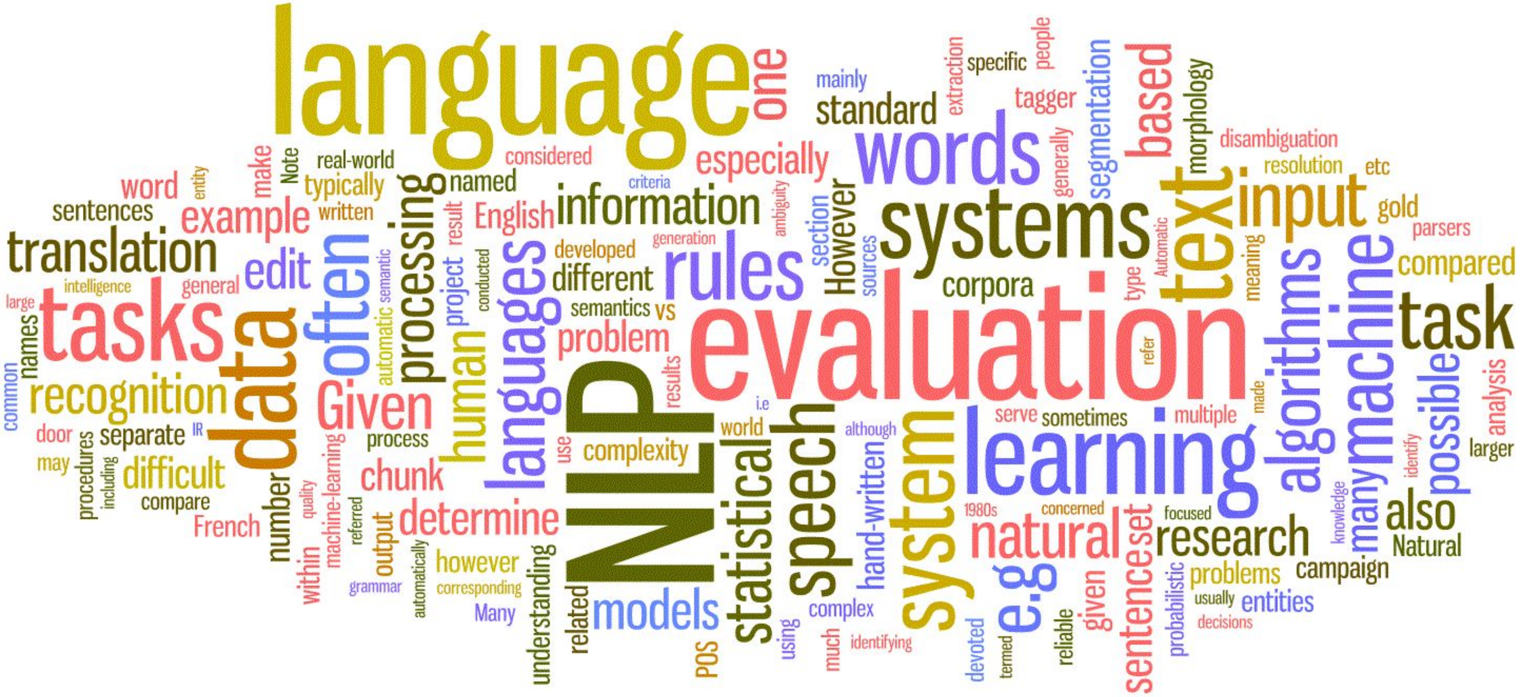


Natural Language Processing

Isac Boström, Rebecka Jacobsson & Britta Thörnblom

Outline

- What Is Natural Language Processing?
- Vector Representation of Words
- Application 1: Machine Translation
- Application 2: Generating Clickbait Headlines
- Summary

What is Natural Language Processing?

Aim: Create computer systems that can understand and output *“natural language”*.

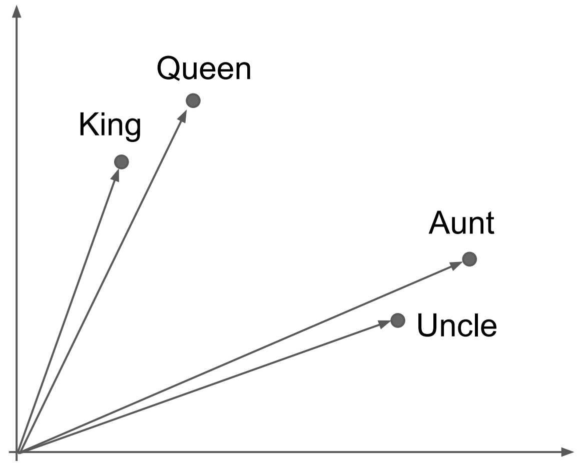
- ➔ So what is a *natural language*?
  - ◆ Language that develops naturally in humans
  - ◆ In practice: Used for communication between humans, e.g English
  - ◆ Opposite: Constructed languages, eg. Python

- ➔ And what is the point?
  - ◆ Path to AI
  - ◆ Simplify communication between humans and machines

- ➔ What are the major subtasks?
  - ◆ Automatic text summarization, classification and translation
  - ◆ Text generation and question-answering
  - ◆ Speech recognition and sentiment analysis

Vector representation

- ➔ How to teach a machine to understand meaning and context of a word?
- ➔ Vector representation of words
- ➔ The cosine distance between the vectors tells us how the respective words are related
- ➔ Neural network

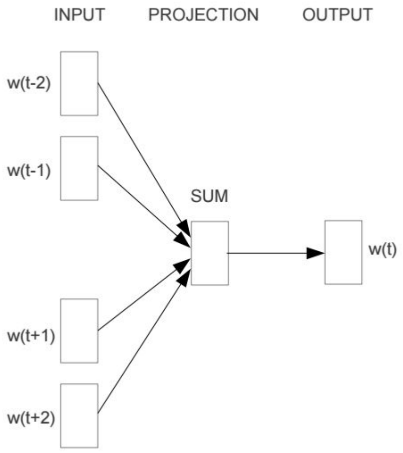


Vector representation - neural networks

Two large data sets.

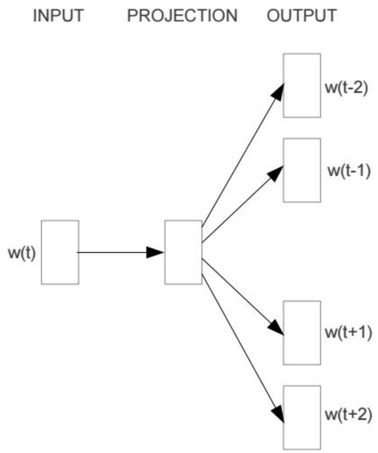
The dog chases the ___.

Either determine a word given a context, or determine a context given a word.



CBOW

(Continuous bag-of-words)



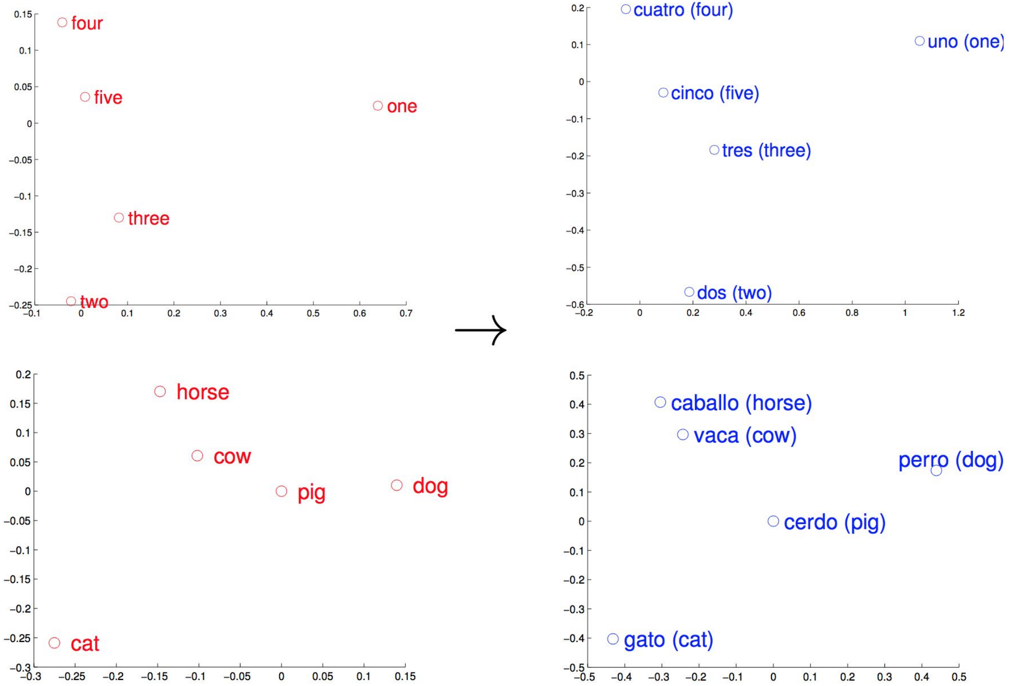
Skip-gram

Vector representation - results

Using simple vector operations:

- ➔ King - Man + Woman = Queen
- ➔ Windows - Microsoft + Google = Android
- ➔ Knee is to Leg as Elbow is to … [Forearm, Arm, Ulna bone]

Machine translation



English

Spanish

Basic idea:
The relations between simple words are similar in most languages

Mikolov et al 2013

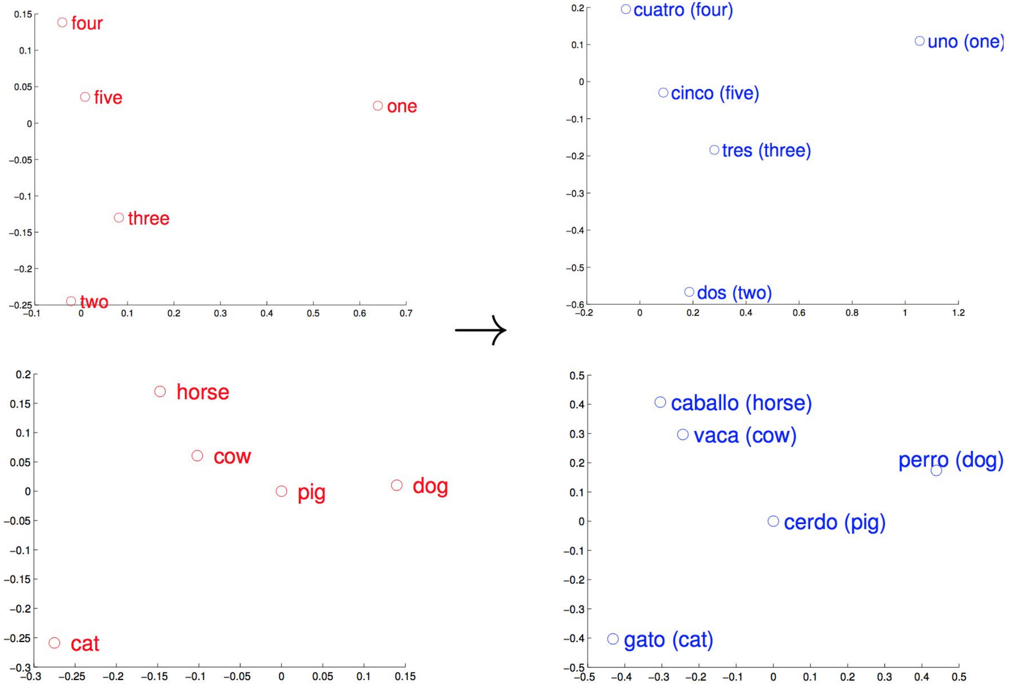
Generating a dictionary: a how-to

1. Construct a word space for each language using large amounts of text
2. Learn the linear projection between the languages

➔ Makes it possible to project a word from one word space to another

➔ Output is the vector of the second word space most similar to the projection

A glimmer of the maths



English

Spanish

Vector representation of the original word

$$\min_{W} \sum_{i=1}^{n} \|Wx_i - z_i\|^2$$

Transformation matrix

Vector in target language space closest to the transformation

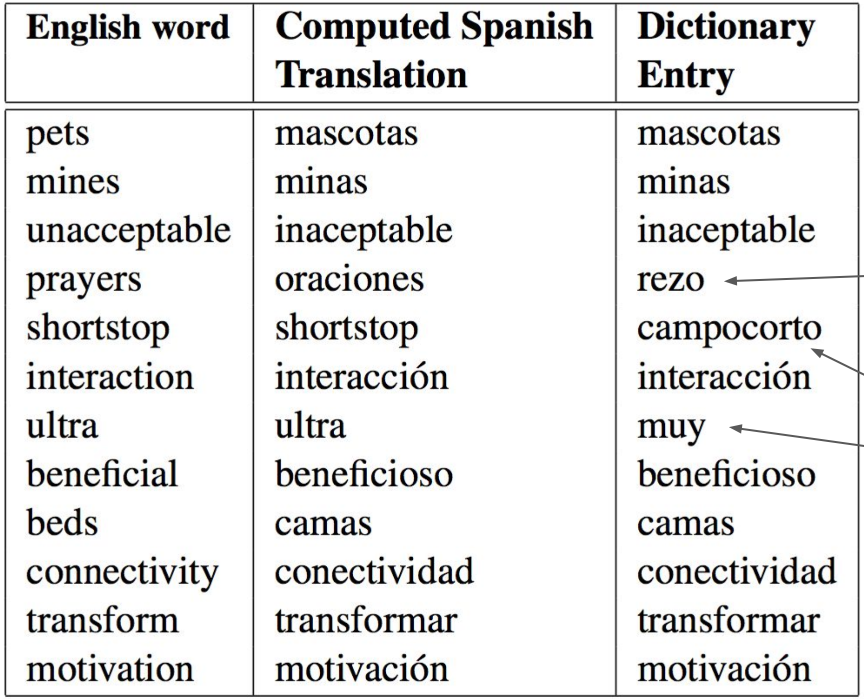
Weeding out misinterpretations

Confidence measure: the cosine between transformed vector and output vector

Translations below a certain threshold are discarded

Some results

Translations from english to spanish with confidence scores exceeding 0.5



English word	Computed Spanish Translation	Dictionary Entry
pets	mascotas	mascotas
mines	minas	minas
unacceptable	inaceptable	inaceptable
prayers	oraciones	rezo
shortstop	shortstop	campocorto
interaction	interacción	interacción
ultra	ultra	muy
beneficial	beneficioso	beneficioso
beds	camas	camas
connectivity	conectividad	conectividad
transform	transformar	transformar
motivation	motivación	motivación

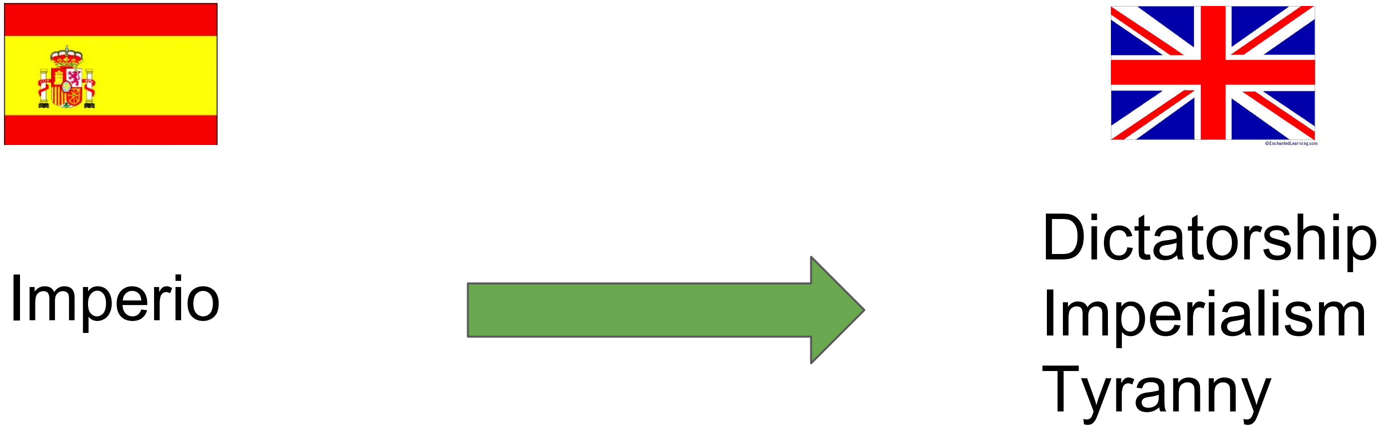
Overall: solid translations

Oraciones = prayers
Rezo = I pray

Loan words, probably more commonly used in their original form

Logical errors

Example:



Imperio

Dictatorship
Imperialism
Tyranny

Correction of existing dictionaries

1. Translate a number of words
2. Compute the confidence measures compared to an existing dictionary
3. Discard those with **high** confidence measures

➔ The remaining translations have little overlap between dictionary and computed translation

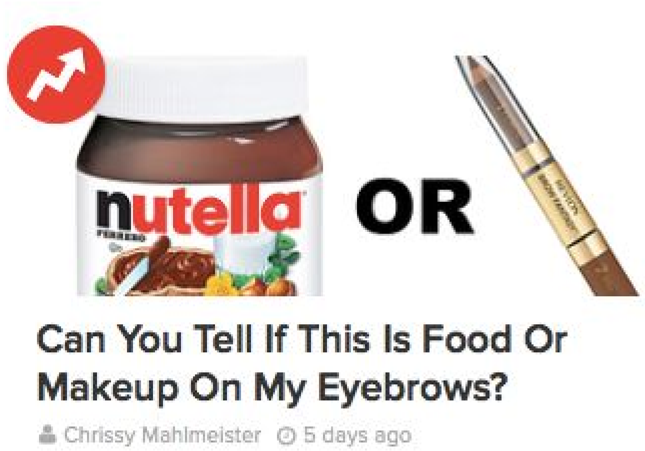
➔ These dictionary entries can be corrected without having to revise the whole dictionary

This Application of Automated Text-Generation Will Shock You. It’s Amazing!

Clickbait: Web content aimed at generating advertising revenue, relying on sensationalist headlines to attract click-throughs and encourage forwarding over online social networks - Wikipedia

Suitable because:

- ➔ Headlines are short
- ➔ Simple language
- ➔ Small syntactic variation
- ➔ Correct grammar not necessary

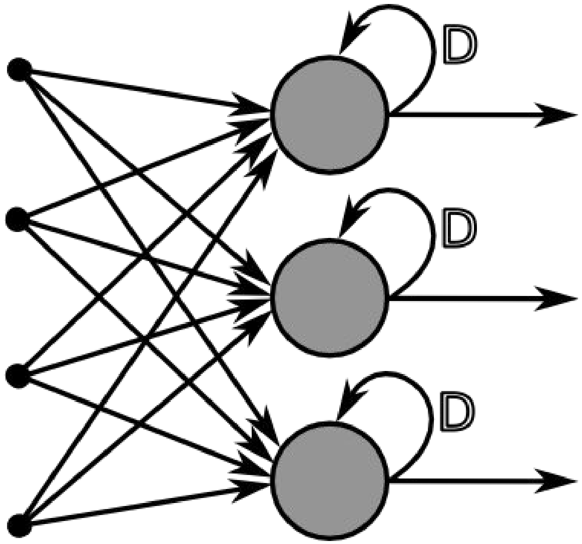


Can You Tell If This Is Food Or Makeup On My Eyebrows?

Chrissy Mahlmeister 5 days ago

Four Facts About Recurrent Neural Networks Every CAS Student Should Know - You Won’t Believe Number Three!

- RNN:s have memory, which makes them suitable for NLP
- Standard training method is back propagation through time (BPTT)
- The contribution to the gradient decreases exponentially with the time lag between inputs
- The long short term memory (LSTM) architecture provides a solution



How The Clickbait Network Was Trained - You Can’t Imagine What Happened Next!

Input

- 2 million headlines from Buzzfeed, Gawker, Jezebel, Huffington Post and Upworthy
- Pre-trained word vector representations

➡

Training

- RNN network with two recurrent LSTM layers

➡ ?

13 Crazy Clickbait Headlines Created by Machines

John McCain Warns Supreme Court To Stand Up For Birth Control Reform

A Tour Of The Future Of Hot Dogs In The United States

Mary J . Williams On Coming Out As A Woman

Miley Cyrus Turns 13

Romney Camp: ‘I Think You Are A Bad President'

How To Use The Screen On The IPhone 3 Music Player

How To Get Your Kids To See The Light

Barack Obama Says …

… It’s Wrong To Talk About Iraq

… GOP Needs To Be Key To New Immigration Policy

… He Is Wrong

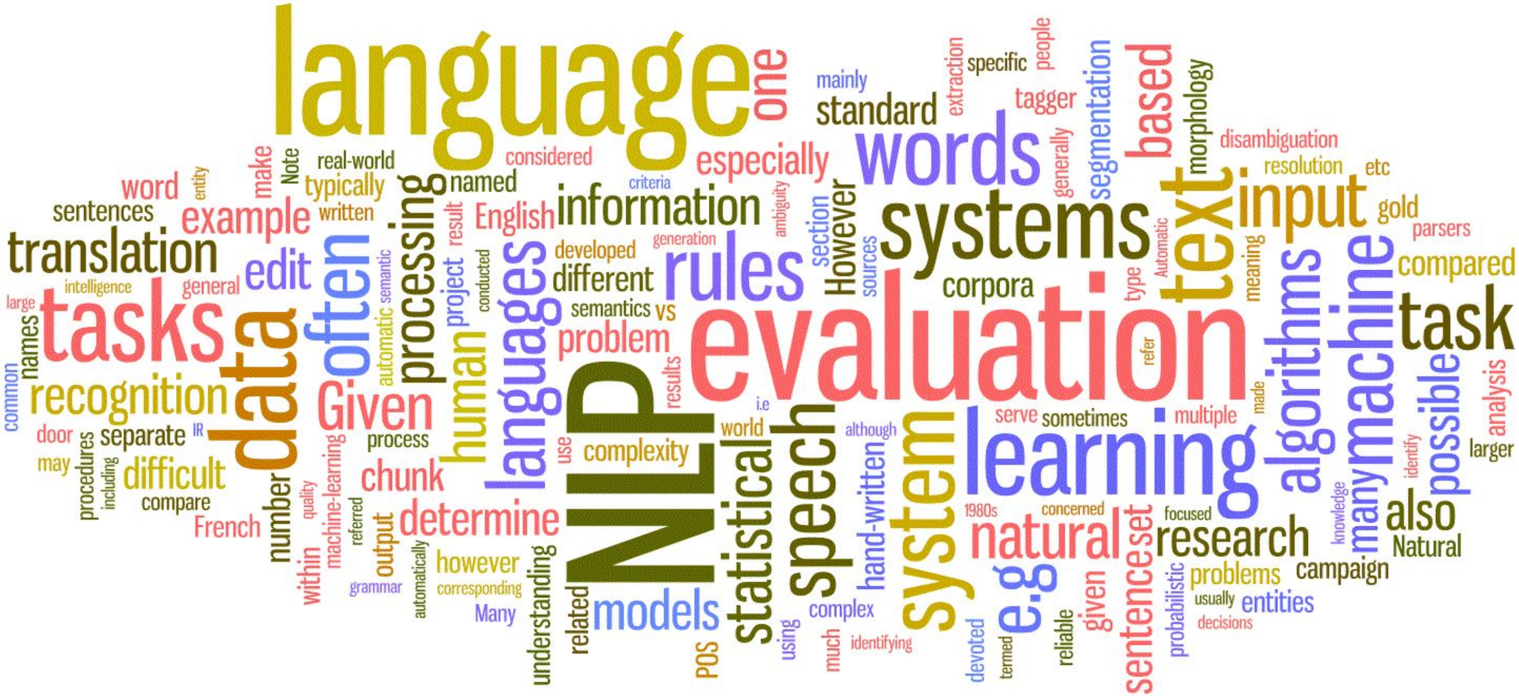
Kim Kardashian Says …

… Kanye West Needs A Break From Her

… She Looks Fake

… She’s Married With A Baby In New Mexico

Summary


language
evaluation
NLP
learning
systems
words
text
rules
data
tasks
languages
translation
speech
system
information
input
task
machine
algorithms
often
processing
statistical
natural
research
human
recognition
Given
determine
models
especially
standard
based
edit
example
problem
different
corpora
possible
also
e.g
sentence
set
one
word
sentences
typically
named
English
considered
real-world
make
Note
entity
written
result
project
developed
semantics
vs
generation
criteria
ambiguity
section
However
sources
mainly
extraction
specific
people
tagger
generally
segmentation
morphology
disambiguation
resolution
etc
gold
parsers
compared
Automatic
type
meaning
intelligence
general
large
automatic
semantic
conducted
names
common
door
separate
IR
may
difficult
procedures
including
compare
French
number
within
quality
machine-learning
referred
output
process
chunk
automatically
grammar
however
corresponding
Many
use
complexity
world
results
i.e
understanding
related
POS
using
although
hand-written
complex
much
identifying
serve
sometimes
multiple
refer
made
1980s
concerned
devoted
termed
reliable
given
probabilistic
focused
usually
problems
decisions
entities
campaign
many
knowledge
identify
Natural
larger
analysis

Further Reading / References

- Glove and Word2Vec
- Mikolov et al, all articles from 2013 but especially
    - Efficient Estimation of Word Representations in Vector Space
    - Exploiting Similarities among Languages for Machine Translation
- Auto-Generating Clickbait With Recurrent Neural Networks
- Examples of Auto-Generated Headlines

Discussion Questions

1. In the presentation we were introduced to a number of erroneous translations caused either by the machine misinterpreting the input or by being unable to translate loan words. Are there other types of text which could be hard to implement NLP on?

2. Can you think of other fields where the principle of vector representation could be applied?

3. In the start of the seminar the presenters mentioned speech recognition as an application of NLP. What are the major difficulties with such an application and does is seem feasible to develop a universal speech recognising algorithm? In which ways could different forms of NLP make human-machine interaction easier?

4. Computer-generated news articles are already commonplace in many fields of journalism. Can you think of any possible issues with having more and more of our information channels generated without human input?