Foundations of Artificial Intelligence

15. Natural Language Processing

Understand, interpret, manipulate, generate human language (text and audio)

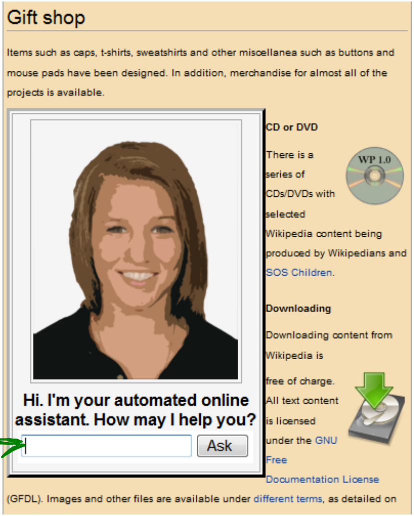
Joschka Boedecker and Wolfram Burgard and
Frank Hutter and Bernhard Nebel and Michael Tangermann

Albert-Ludwigs-Universität Freiburg

July 17, 2019

Contents

Example: Automated Online Assistant



Source: Wikicommons/Bemidji State University

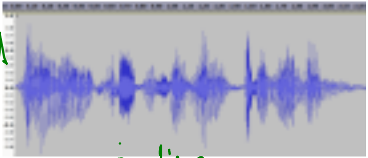
Lecture Overview

1. Motivation, NLP Tasks
2. Learning Representations
3. Sequence-to-Sequence Deep Learning

Natural Language Processing (NLP)

text

audio



time

Credits: slide by Torbjoern Lager; (audio: own)

- The language of humans is represented as text or audio data. The field of NLP creates interfaces between human language and computers.
- Goal: automatic processing of large amounts of human language data.

Examples of NLP Tasks and Applications

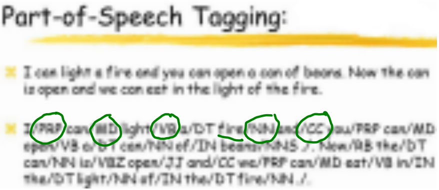
- word stemming
- word segmentation, sentence segmentation
- text classification
- sentiment analysis (polarity, emotions, ..)
- topic recognition
- automatic summarization
- machine translation (text-to-text)

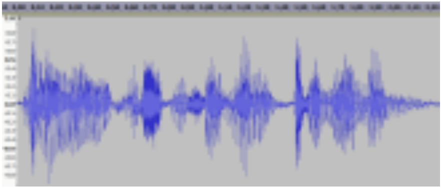
text-based

- speaker identification
- speech segmentation (into sentences, words)
- speech recognition (i.e. speech-to-text)
- natural language understanding
- text-to-speech
- text and spoken dialog systems (chatbots)

audio-based

From Rules to Probabilistic Models to Machine Learning





Sources: Slide by Torbjoern Lager; (Anthony, 2013)

Traditional rule-based approaches and (to a lesser degree) probabilistic NLP models faced limitations, as

- human don't stick to rules, commit errors.
- language evolves: rules are neither strict nor fixed.
- labels (e.g. tagged text or audio) were required.

Machine translation was extremely challenging due to shortage of multilingual textual corpora for model training.

From Rules to Probabilistic Models to Machine Learning

Machine learning entering the NLP field:

- Since late 1980's: increased data availability (WWW)
- Since 2010's: huge data, computing power $\rightarrow$ unsupervised representation learning, deep architectures for many NLP tasks.

Lecture Overview

1 Motivation, NLP Tasks

2 Learning Representations

3 Sequence-to-Sequence Deep Learning

Learning a Word Embedding

(https://colah.github.io/posts/2014-07-NLP-RNNs-Representation)

A word embedding W is a function

$$W : \text{words} \rightarrow \mathbb{R}^n$$

200 dim

which maps words of some language to a high-dimensional vector space (e.g. 200 dimensions).

Examples:

$$W(\text{"cat"}) = (0.2, \text{-}0.4, 0.7, \ldots)$$
$$W(\text{"mat"}) = (0.0, 0.6, \text{-}0.1, \ldots)$$

Mapping function W should be realized by a look-up table or by a **neural network** such that:

- representations in $\mathbb{R}^n$ of related words have a short distance
- representations in $\mathbb{R}^n$ of unrelated words have a large distance

How can we learn a good representation / word embedding function W?

Representation Training

A word embedding function W can be trained using different tasks, that require the network to discriminate related from unrelated words.

Can you think of such a training task? Please discuss with your neighbors!

Representation Training

A word embedding function W can be trained using different tasks, that require the network to discriminate related from unrelated words.

Can you think of such a training task? Please discuss with your neighbors!

Representation Training

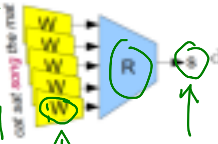
A word embedding function W can be trained using different tasks, that require the network to discriminate related from unrelated words.

Example task: predict, if a 5-gram (sequence of five words) is valid or not. Training data contains valid and slightly modified, invalid 5-grams:

$$R(W(\text{"cat"}), W(\text{"sat"}), W(\text{"on"}), W(\text{"the"}), W(\text{"mat"})) = 1$$
$$R(W(\text{"cat"}), W(\text{"sat"}), W(\text{"song"}), W(\text{"the"}), W(\text{"mat"})) = 0$$
$$\ldots$$

classif. task

Train the combination of embedding function W and classification module R:



Modular Network to determine if a 5-gram is 'valid' (From Bottou (2011))

While we may not be interested in the trained module R, the learned word embedding W is very valuable!

Visualizing the Word Embedding

Let's look at a projection from $\mathbb{R}^n \to \mathbb{R}^2$ obtained by tSNE:

Visualizing the Word Embedding

Let's look at a projection from $\mathbb{R}^n \to \mathbb{R}^2$ obtained by tSNE:

t-SNE visualizations of word embeddings. Left: Number Region; Right: Jobs Region. From Turian *et al.* (2010)

Sanity Check: Word Similarities in $\mathbb{R}^n$?

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

What words have embeddings closest to a given word? From Collobert *et al.* (2011)

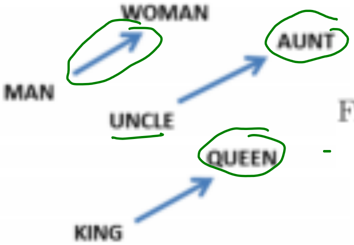
Powerful Byproducts of the Learned Embedding W

Embedding allows to work not only with synonyms, but also with other words of the same category:

- "the cat is black" $\rightarrow$ "the cat is white"
- "in the zoo I saw an elephant" $\rightarrow$ "in the zoo I saw a lion"

In the embedding space, systematic shifts can be observed for analogies:

W(Woman) − W(Man)



From Mikolov *et al.* (2013a)

The embedding space may provide dimensions for gender, singular-plural etc.!

Observed Relationship Pairs in the Learned Embedding W

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Relationship pairs in a word embedding. From Mikolov *et al.* (2013b).

Word Embeddings Available for Your Projects

Various embedding models / strategies have been proposed:

- Word2vec (Tomas Mikolov et al., 2013)
- GloVe (Pennington et al., 2014)
- fastText library (released by Facebook by group around Tomas Mikolov)
- ELMo (Matthew Peters et al., 2018)
- ULMFit (by fast.ai founder Jeremy Howard and Sebastian Ruder)
- BERT (by Google)
- ...

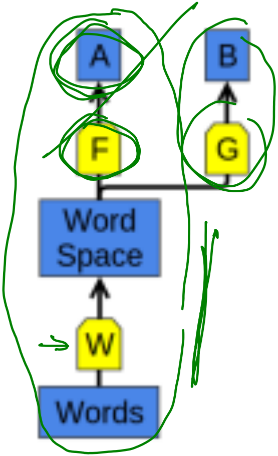
(Pre-trained models are available for download)

Word Embeddings: the Secret Sauce for NLP Projects

Shared representations — re-use a pre-trained embedding for other tasks!

Using ELMo embeddings improved six state-of-the-art NLP models for:

- Question answering
- Textual entailment (inference)
- Semantic role labeling
  ("Who did what to whom?")
- Coreference resolution
  (clustering mentions of the same entity)
- Sentiment analysis
- Named entity extraction



W and F learn to perform task A. Later, G can learn to perform B based on W.

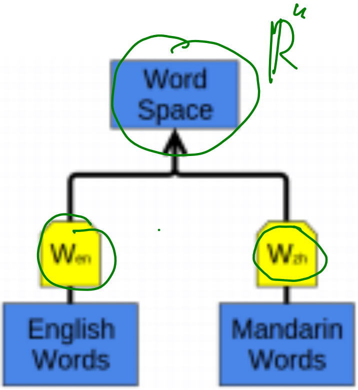
Can Neural Representation Learning Support **Machine Translation**?

Can you think of a training strategy to translate from Mandarin to English and back? Please discuss with your neighbors!

Can Neural Representation Learning Support **Machine Translation**?

Can you think of a training strategy to translate from Mandarin to English and back? Please discuss with your neighbors!

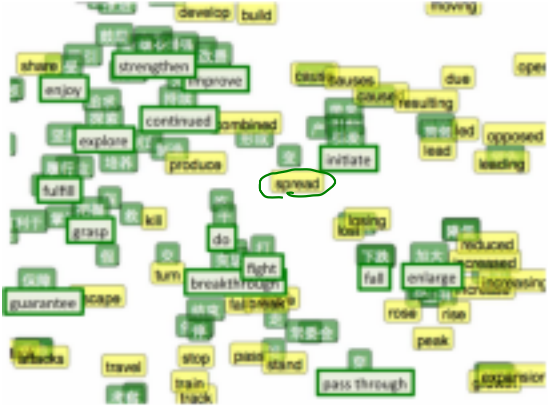
Bilingual Word Embedding



Idea: train two embeddings in parallel such, that corresponding words are projected to close-by positions in the word space.

Visualizing the Word Embedding

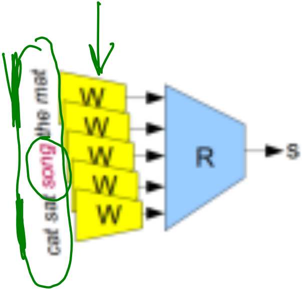
Let's again look at a tSNE projection $\mathbb{R}^n \to \mathbb{R}^2$:



t-SNE visualization of the bilingual word embedding. Green is Chinese, Yellow is English. (Socher *et al.* (2013a))

Lecture Overview

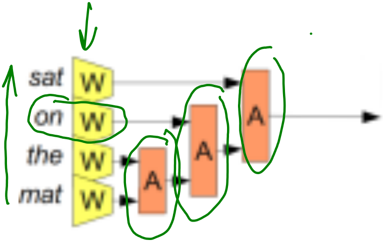
- So far, the network has learned to deal with a **fixed number of input words** only.



Association Modules

RNN

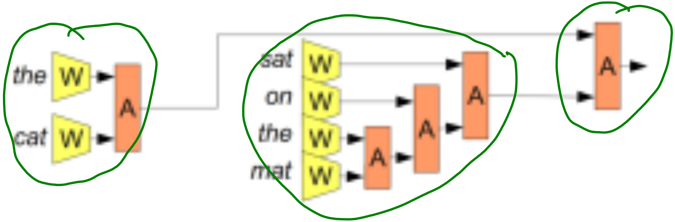
- So far, the network has learned to deal with a **fixed number of input words** only.
- Limitation can be overcome by adding **association modules**, which can combine two word and phrase representations and merge them



(From Bottou (2011))

Association Modules

- So far, the network has learned to deal with a **fixed number of input words** only.
- Limitation can be overcome by adding **association modules**, which can combine two word and phrase representations and merge them
- Using associations, whole sentences can be represented!



(From Bottou (2011))

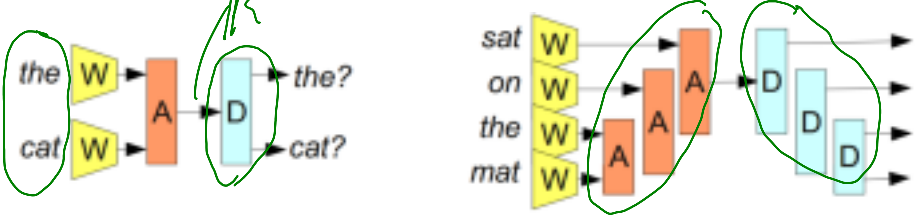
From Representations to the Translation of Texts

Conceptually, we could now use this concept to find the embedding of a word or sentence of the source language and look up the closest embedding of the target language.

What is missing to realize a translation?

For translations, we also need disassociation modules!
(encoder — decoder principle)

encoding decode



(From Bottou (2011))

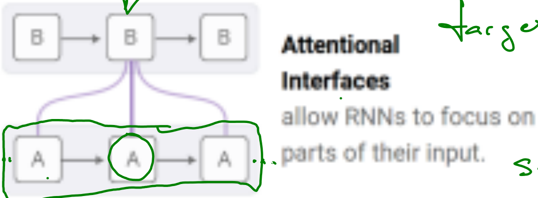
Sequence-to-Sequence Neural Machine Translation

Ground-breaking new approach by Bahdanau, Cho and Bengio (2014 ArXiv, 2015 ICML)

- Shift through the input word sequence
- Learn to encode and to decode using recurrent neural networks (RNN)
- Learn to align input and output word sequences
- Take context into account by learning the importance of neigboring words → **attention mechanism**.

decode

target language



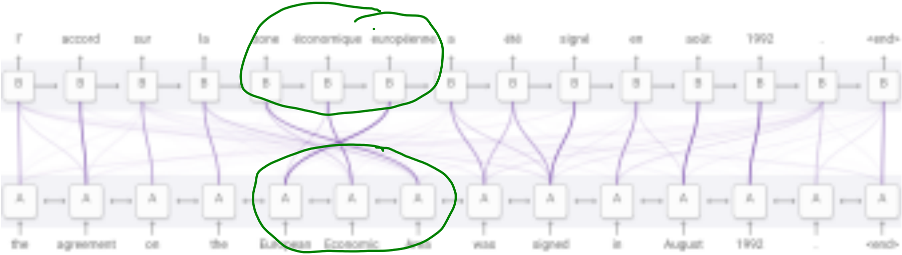
encoding →

source L.

Credits: (Olah & Carter, 2016) have adapted this figure based on (Bahdanau et al., 2014)

Ground-breaking new approach by Bahdanau, Cho and Bengio (2014 ArXiv, 2015 ICML)

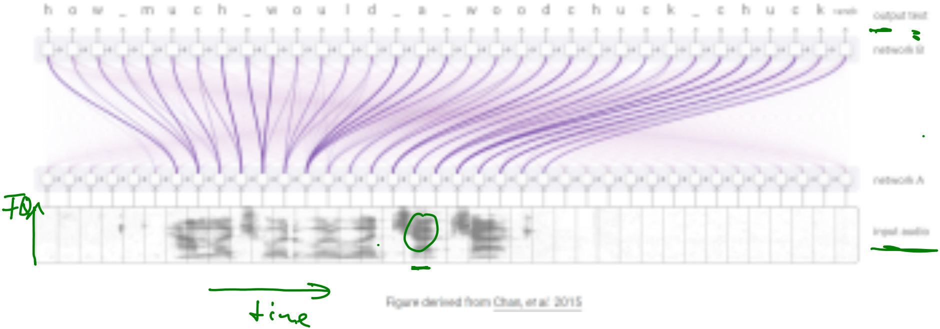
- Shift through the input word sequence
- Learn to encode and to decode using recurrent neural networks (RNN)
- Learn to align input and output word sequences
- Take context into account by learning the importance of neigboring words → **attention mechanism**.



Credits: (Olah & Carter, 2016) have adapted this figure based on (Bahdanau et al., 2014)

Sequence-to-Sequence Neural Voice Recognition

- Similar principle, but voice/speech input

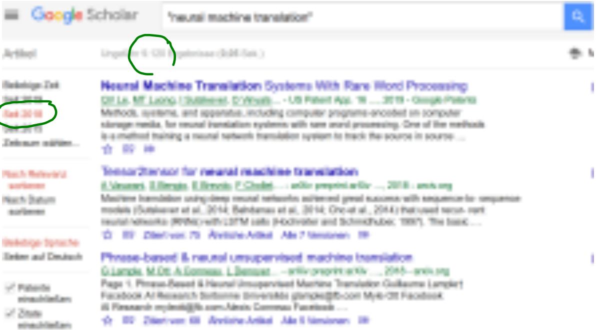


Credits: (Olah & Carter, 2016) have adapted this figure based on (Chan et al., 2015)

Success Story of Attention-based Neural Machine Translation

Neural machine translation requires big data sets but has advantages:

- Overall model can be learned end-to-end
- No need to integrate modules for feature extraction, database, grammar rules etc. in a complicated system



Summary

- Natural language processing spans a wide range of problems and applications.
- NLP is a rapidly growing field due to availability of huge data sets.
- NLP techniques is part of many products already.
- Field is moving more and more to neural networks, which provide NLP building blocks like end-to-end learning, representation learning, sequence-to-sequence, ...