

Natural Language Processing Research in Low-Resourced Languages

Cristina España-Bonet
DFKI GmbH

Making NLP Work in Africa
Teams, everywhere on the Earth with
internet
3rd July 2020

Outline

- 1 Natural Language Processing Today
- 2 Dense Semantic Representations, Embeddings
- 3 Low-Resourced Languages
- 4 Transfer Learning in NLP

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

POS Tagging

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

POS Tagging

Parsing

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

Sentiment Analysis

POS Tagging

Parsing

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

Sentiment Analysis

POS Tagging

Semantic Role Labeling

Parsing

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

Sentiment Analysis

POS Tagging

Question Answering

Semantic Role Labeling

Parsing

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

Sentiment Analysis

POS Tagging

Question Answering

Semantic Role Labeling

Machine Translation

Parsing

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

Sentiment Analysis

POS Tagging

Question Answering

Semantic Role Labeling

Dialogue

Machine Translation

Parsing

Natural Language Processing Today

NLP Tasks

Named-Entity Recognition

Sentiment Analysis

POS Tagging

Question Answering

Semantic Role Labeling

Dialogue

Text Summarisation

Machine Translation

Parsing

Natural Language Processing Today

NLP Tasks

Seq2Tag

Named-Entity Recognition

Sentiment Analysis

POS Tagging

Question Answering

Semantic Role Labeling

Seq2Seq

Dialogue

Text Summarisation

Machine Translation

Parsing

Natural Language Processing Today

Seq2Seq: a single NN for Several Tasks

The screenshot shows a web browser displaying the fairseq documentation. The browser's address bar shows the URL <https://fairseq.readthedocs.io/en/latest/>. The page has a blue header with the 'fairseq' logo and the word 'latest'. A search bar is present. On the left, a dark sidebar lists navigation links under 'GETTING STARTED' and 'EXTENDING FAIRSEQ'. The main content area has a breadcrumb 'Docs » fairseq documentation' and a link to 'Edit on GitHub'. The title 'fairseq documentation' is prominently displayed. Below it, a paragraph describes Fairseq as a sequence modeling toolkit written in PyTorch, used for training custom models for translation, summarization, language modeling, and other text generation tasks. A 'Getting Started' section follows with a bulleted list of links.

fairseq
latest

Search docs

GETTING STARTED

- Evaluating Pre-trained Models
- Training a New Model
- Advanced Training Options
- Command-line Tools

EXTENDING FAIRSEQ

- Overview
- Tutorial: Simple LSTM
- Tutorial: Classifying Names with a Character Level RNN

Read the Docs v: latest ▼

Docs » fairseq documentation [Edit on GitHub](#)

fairseq documentation

Fairseq is a sequence modeling toolkit written in [PyTorch](#) that allows researchers and developers to train custom models for [translation](#), [summarization](#), [language modeling](#) and [other text generation tasks](#).

Getting Started

- [Evaluating Pre-trained Models](#)
- [Training a New Model](#)
- [Advanced Training Options](#)
- [Command-line Tools](#)

Natural Language Processing Today

Infinite Variations of the NNs

README.md



Transformers

build passing

license Apache-2.0

website online

release v2.11.0

State-of-the-art Natural Language Processing for PyTorch and TensorFlow 2.0

🤗 Transformers (formerly known as `pytorch-transformers` and `pytorch-pretrained-bert`) provides state-of-the-art general-purpose architectures (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, T5, CTRL...) for Natural Language Understanding (NLU) and Natural Language Generation (NLG) with over thousands of pretrained models in 100+ languages and deep interoperability between PyTorch & TensorFlow 2.0.

Outline

- 1 Natural Language Processing Today
- 2 Dense Semantic Representations, Embeddings
- 3 Low-Resourced Languages
- 4 Transfer Learning in NLP

Dense Semantic Representations, Embeddings

A World where Words are Numbers

Basketball = (0.101159, 0.550446, 0.543801, -0.973852, -0.680835, 0.417193, -0.247181, 0.209725, -1.136055, -0.059531, -0.401640, 0.171540, 0.925121, -0.143815, 0.781714, -1.482425, 0.347008, -0.112342, 0.442418, -1.020457, -0.071752, 1.873548, -0.222886, -0.729569, -0.830224, -0.868407, 0.203496, 0.469911, -0.191363, 0.565102, 0.687738, 0.480823, 0.842358, -0.173656, -0.265585, 0.685740, 0.488047, -0.359772, -0.576064, -0.802884, 0.081554, 0.046882, -0.861532, -0.461855, 0.613098, -1.534642, -0.884534, 0.207728, 1.396512, -0.242900, -0.383959, 0.570844, -0.703350, -1.368813, -1.008194, 1.534660, 0.171693, 0.640925, -0.233116, 0.324685, 0.483171, 0.337947, -0.963290, -0.400558, 0.830977, 0.913474, 0.251693, -0.589420, -0.299622, 1.047515, -0.266679, -1.247186, 1.087610, -0.549028, 1.600710, -1.538516, -1.703301, -1.393499, -0.894448, 0.717204, 0.105767, -0.189234, -0.615609, -0.658315, 0.051877, 0.014180, -0.791282, 0.150424, 1.343751, -0.464859, 0.871426, 1.542864, -1.202150, -0.767113, -1.734738, 0.073633, -1.012583, 0.747787, 0.476070, -0.454807, 0.642685, -0.854152, -0.071798, 0.233724, 0.712329, -0.097752, -0.531132, 0.323271, -0.447342, 0.657913, 1.199492, -0.107360, -0.154234, -1.131168, 1.354793, 1.721385, -0.240023, 0.655765, -0.217006, -0.801722, 0.553369, 0.213377, 0.323267, -1.516051, 2.106244, -0.134282, 0.742155, 0.426344, 0.197991, -0.806768, 0.372546, -0.160200, -1.552847, -0.286178, -0.707796, 0.527352, -0.259658, 0.230387, 0.105294, -0.194481, 0.301772, -1.022163, 0.557191, 1.096709, 0.058422, -1.036384, 0.353412, -0.623097, -0.689515, 0.091472, 0.783885, 0.184088, -0.367950, 0.952462, 0.183704, 0.677562, 0.293917, -0.214309, -0.487794, 0.934296, 0.311513, 0.286514, -0.085511, 0.777691, 1.232603, -0.309367, -0.225086, 0.005091, -0.099195, -0.293117, 1.305563, 0.595816, 0.950316, 0.568706, -0.561446, 0.911634, -0.383941, 0.758054, -0.197820, 0.506777, -0.290767, -0.356727, 1.229474, -0.156489, -0.782741, -0.210163, -0.029169, 0.602664, 0.418375, 0.148975, -0.761796, 1.322690, -0.173410, 0.204111, -1.344531, 1.081905, -0.660543, -0.225615, -0.444753, -0.929671, 0.054136, 0.052031, -0.164926, 0.159312, -1.316333, 0.837011, -1.290353, 0.958403, 1.247478, 0.442009, 0.455497, -1.856268, -0.358823, -0.230839, -0.206271, 0.227012, -0.454163, 0.747798, -1.252855, 1.436849, -0.427915, -0.810428, -0.628144, -0.288458, 0.087355, 0.356739, 0.153036, 0.516594, -0.504978, 0.814432, 1.052940, 1.094526, -0.219595, 0.722178, 0.267325, -0.087458, -1.270262, -0.039461, 0.991926, -0.112005, -0.009605, 0.149920, 0.164717, 0.280475, 0.966384, 0.327598, 0.189590, -0.208946, 0.838261, 0.051847, -0.277932, -0.788527, -0.768702, -1.688721, 0.388215, 0.170153, -0.555723, -0.529565, -0.528982, -0.659930, 0.588041, -0.368195, -0.850188, -0.004996, 0.925476, 1.046587, -0.731761, 0.519435, 0.193188, -0.709557, 0.123329, -0.454316, 1.885830, -0.201841, -0.728933, -0.953455, -0.205837, -0.724068, 0.120158, 1.765389, -0.192159, 1.062490, -0.002634, 0.125790, -0.846565, 0.548899, -1.062821, -2.146826, 0.134681, 0.570950, 0.851783, 0.436544, 0.688986, 1.229008, 1.435449, 0.118766, -0.132411, 2.527890, 0.778142)

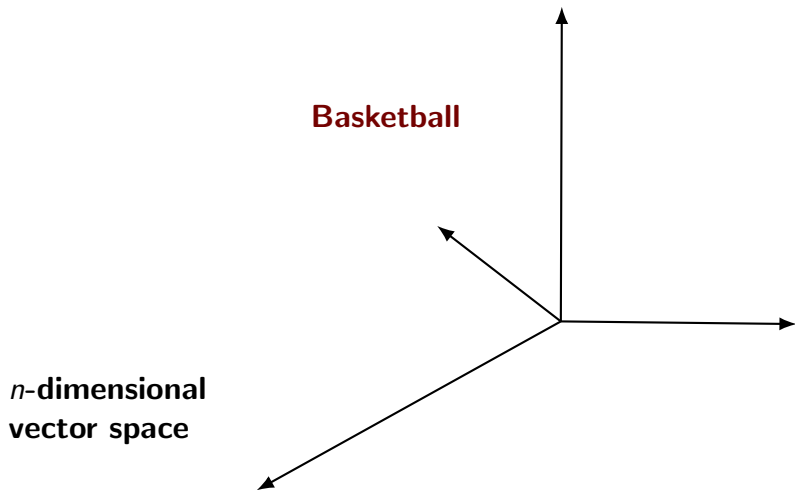
Dense Semantic Representations, Embeddings

Word Vector Space

Basketball

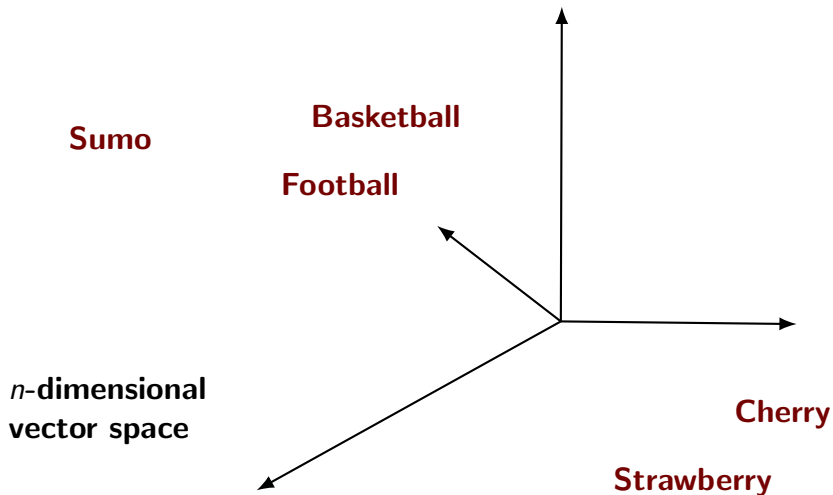
Dense Semantic Representations, Embeddings

Word Vector Space



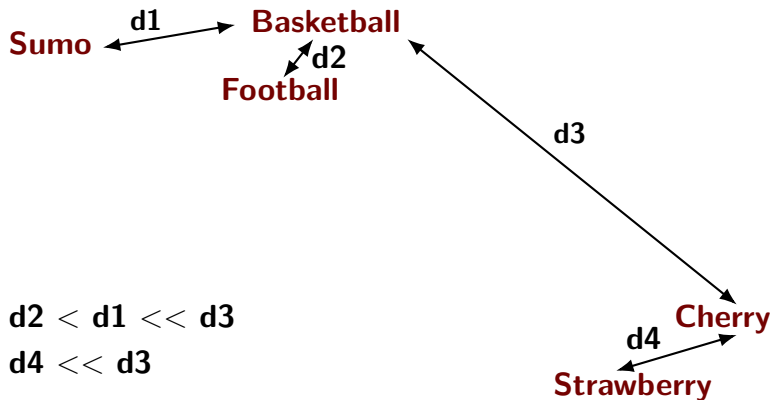
Dense Semantic Representations, Embeddings

Word Vector Space



Dense Semantic Representations, Embeddings

Distances in Space Account for Semantics



Dense Semantic Representations, Embeddings

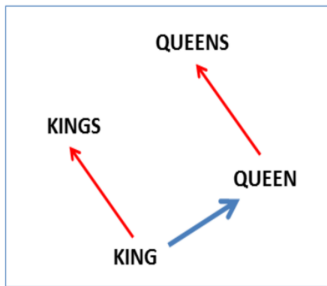
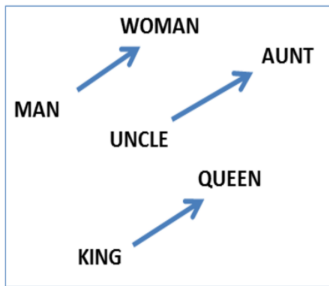
Distances in Space Account for Semantics

Would that allow language understanding?

Dense Semantic Representations, Embeddings

Word Embeddings

King - Man + Woman = Queen

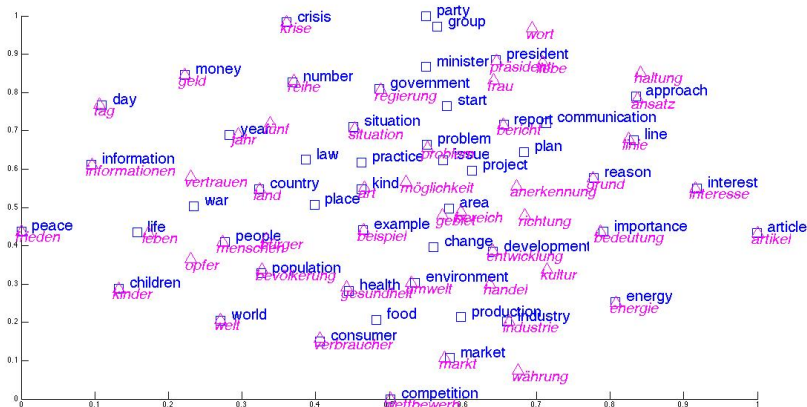


(Mikolov et al., NAACL HLT, 2013)

Dense Semantic Representations, Embeddings

Multilinguality

(Luong, Pham & Manning, NAACL, 2015)

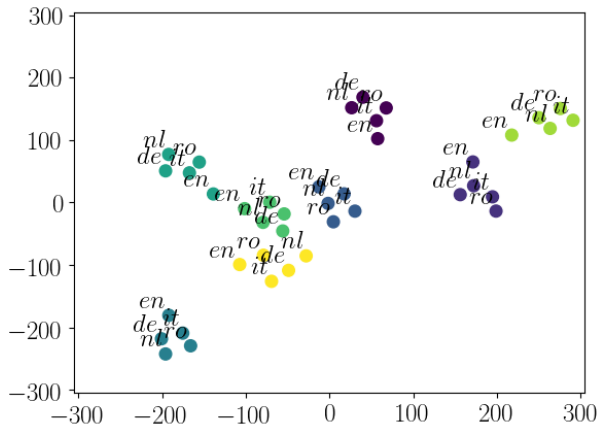


*Barnes-Hut-SNE visualisation of bilingual
embeddings German/English*

Dense Semantic Representations, Embeddings

Multilinguality & Sentence Embeddings

(España-Bonet & van Genabith, 2018)



ML-NMT $\{de, en, nl, it, ro\} \rightarrow \{de, en, nl, it, ro\}$ with TED talks

Dense Semantic Representations, Embeddings

Contextual Embeddings

Question 5: What's the difference between Word2Vec and BERT?

11 responses

I don't know.

Bert is attention model

I don't know, but I checked on google and BERT considers the context while Word2Vec doesn't

Dense Semantic Representations, Embeddings

Word/Contextual Embeddings

Word vectors and contextual word vectors are used differently:

- **Word vectors:** dictionary look-up of words and their corresponding vectors, they are static entities
 - Good to initialise input word embeddings in several NLP tasks

Dense Semantic Representations, Embeddings

Word/Contextual Embeddings

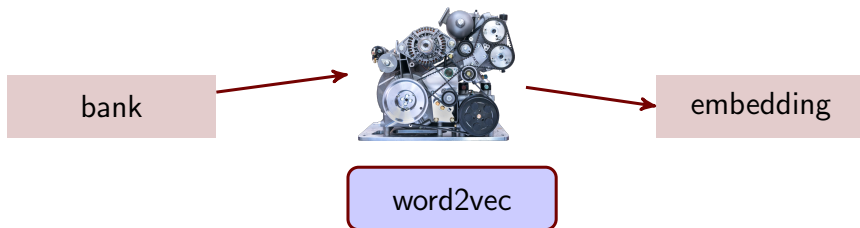
Word vectors and contextual word vectors are used differently:

- **Word vectors:** dictionary look-up of words and their corresponding vectors, they are static entities
 - Good to initialise input word embeddings in several NLP tasks
- **Contextual word vectors:** vectors on-the-fly by passing text through a deep learning model, they would only be static if we could generate all sentences in a language!
 - Good for transfer learning into several NLP tasks (SotA in lots of tasks!)

Dense Semantic Representations, Embeddings

Word/Contextual Embedding

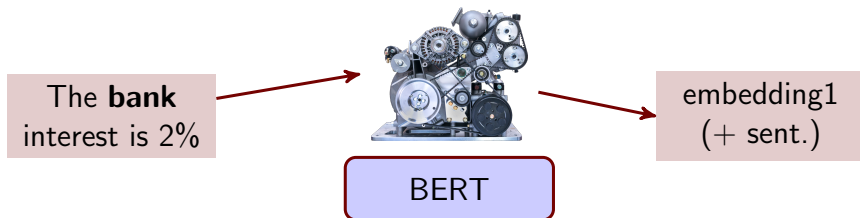
What's the representation for **bank**?



Dense Semantic Representations, Embeddings

Word/Contextual Embeddings

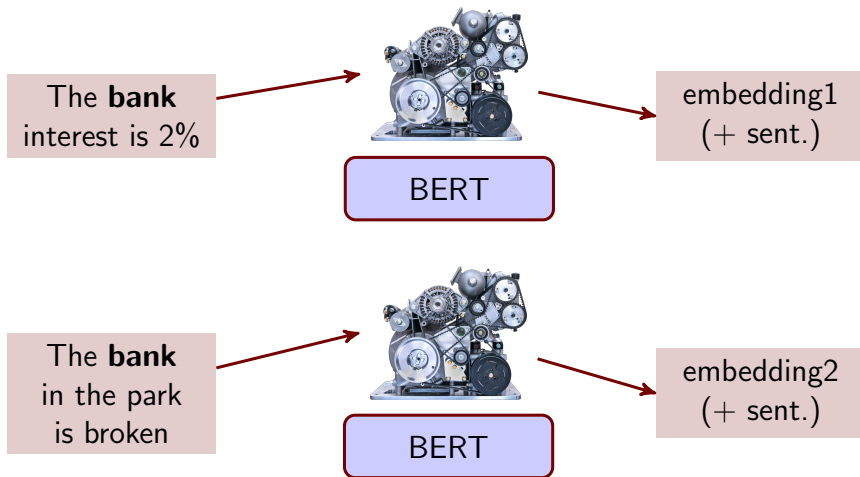
What's the representation for **bank**?



Dense Semantic Representations, Embeddings

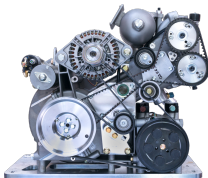
Word/Contextual Embeddings

What's the representation for **bank**?



Dense Semantic Representations, Embeddings

Word2vec and BERT, the neural network behind



word2vec

Feedforward network trained to predict the current word from a window of surrounding context words (CBOW architecture)

BERT

Transformer network trained to predict the next sentence and a masked language model

Dense Semantic Representations, Embeddings

Contextual Embeddings: BERT and Family

[**https://github.com/huggingface/transformers**](https://github.com/huggingface/transformers)

The list is too long for a slide!

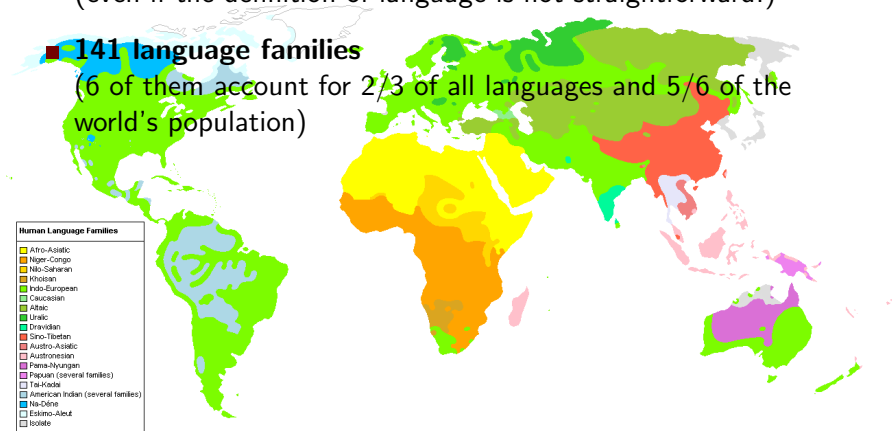
Outline

- 1 Natural Language Processing Today
- 2 Dense Semantic Representations, Embeddings
- 3 Low-Resourced Languages
- 4 Transfer Learning in NLP

Low-Resourced Languages

Some Numbers

- There are more than **7000 languages**
(even if the definition of language is not straightforward!)
- **141 language families**
(6 of them account for 2/3 of all languages and 5/6 of the world's population)



Low-Resourced Languages

Resources and Tasks

- 45% of the content in internet is in English and Chinese (Statista statistics)
- 6950 languages are low-resourced?
- Encouraging data collection
- Transfer learning
- 6000 languages are low-resourced?

Transfer Learning in NLP

Types

Transfer learning in **machine learning**:

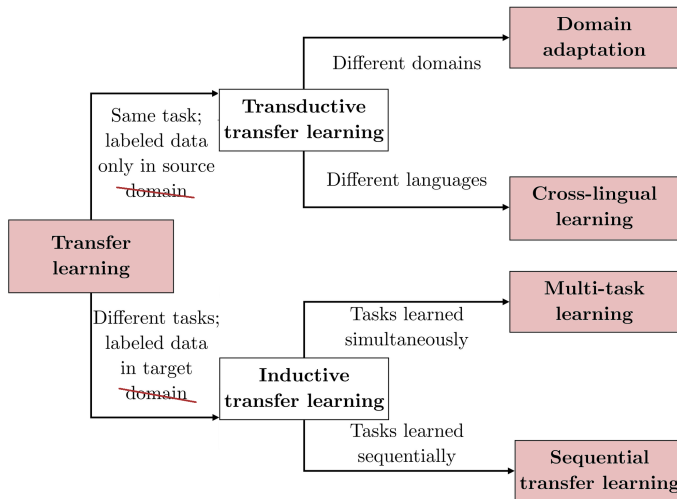
Apply the models obtained solving one problem to a different but related problem

Means for **NLP**:

- Domain adaptation
- Cross-lingual learning
- Multi-task learning
- Sequential transfer learning

Transfer Learning in NLP

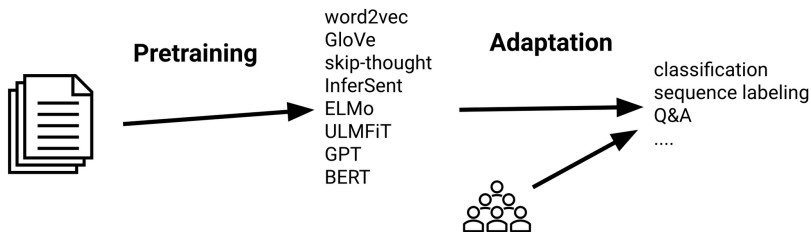
Types



Transfer Learning in NLP

BERT's Success

Sequential task transfer learning:



Transfer Learning in NLP


Use Case: African Languages

Cross-lingual transfer learning:

- Transfer from multilingual models into the desired language
- Transfer from a huge resource (in English?) into the desired language
- Specific examples in the next talk by Jesujoba Alabi

Thanks! And...

wait!



Questions?

Natural Language Processing Research in Low-Resourced Languages

Cristina España-Bonet
DFKI GmbH

Making NLP Work in Africa
Teams, everywhere on the Earth with
internet
3rd July 2020