
**Budapesti Mûszaki és Gazdaságtudományi Egyetem**
Villamosmérnöki és Informatikai Kar
Department of Automation and Applied Informatics

# Universal embedding

Diplomaterv

*Készítette*
Eszter Iklódi

*Konzulens*
Gábor Recski

March 15, 2018

# Contents

## HALLGATÓI NYILATKOZAT

Alulírott *Eszter Iklódi*, szigorló hallgató kijelentem, hogy ezt a szakdolgozatot/ diplomatervet (nem kívánt törlendõ) meg nem engedett segítség nélkül, saját magam készítettem, csak a megadott forrásokat (szakirodalom, eszközök stb.) használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelmûen, a forrás megadásával megjelöltem.

Hozzájárulok, hogy a jelen munkám alapadatait (szerzõ(k), cím, angol és magyar nyelvû tartalmi kivonat, készítés éve, konzulens(ek) neve) a BME VIK nyilvánosan hozzáférhetõ elektronikus formában, a munka teljes szövegét pedig az egyetem belsõ hálózatán keresztül (vagy autentikált felhasználók számára) közzétegye. Kijelentem, hogy a benyújtott munka és annak elektronikus verziója megegyezik. Dékáni engedéllyel titkosított diplomatervek esetén a dolgozat szövege csak 3 év eltelte után válik hozzáférhetõvé.

Budapest, March 15, 2018

*Eszter Iklódi*
hallgató

# Kivonat

Mindennapi életünkben egyre fontosabb szerepet tölt be a természetes nyelv számítógép segítségével történő feldolgozása. Digtitalizált világunkban egyre inkább alapkövetelmény, hogy a gép és ember közötti kommunikáció természetes nyelven történjen. Ennek a megvalósításához elengedhetetlen az emberi nyelv szemantikai értelmezése.

Manapság a state-of-the-art rendszerekben a szavak szemantikai reprezentációja sokdimenziós vektorokkal, word embedingek-kel történik. Diplomaterv munkámban már feltanított word embeddingek-hez keresek olyan fordítási mátrixokat, amelyek képesek egy adott nyelvű word embedding univerzális térbe történő leképzésére.

Az így nyert fordítási mátrixokat különböző feladatokon értékelem ki. (eredmények 2 mondatban)

# Abstract

Computer-driven natural language processing plays an increasingly important role in our everyday life. In our digital world using natural language for machine-human communication is becoming more and more a basic requirement. In order to achieve this goal it is of crucial importance to analyze human languages semantically.

Nowadays state-of-the-art systems represent word meaning with high dimensional vectors, i.e. with word embeddings. In my thesis work I am searching for translation matrices to already trained word embeddings, such that the translation matrices will be able to map these embeddings into a universal space.

These translatin matrices will be evaluated on different tasks. (2 sentences about results)

# Chapter 1

# Introduction

[1]

## 1.1 NLP

A számítástechnika az utóbbi évtizedekben jelentős fellődést mutatott a számítógépen nyelvészet területén. Habár jó néhány nyelvtechnológiai feladatot már javarászt megoldottnak tekinthetünk (spam detektálás, POS-tagging, Named Entity Recognition), és számos további feladaton értünk el jó eredményeket (sentiment analysis, word sense disambiguation, szintaktikai elemzés, gépi fordítás...), azonban még mindig hatalmas mennyiségű megoldatlan feladattal állunk szemben (kérdés megválaszolás, dialógus lebonyolítása, összefoglalás, parafrázisok...).

### 1.1.1 What is NLP, application

### 1.1.2 Common NLP tasks (POS tagging, sentiment analysis, Machine Translation ...)

### 1.1.3 Motivation for NLP research

## 1.2 Universal Embeddings ( 1.5 page)

Summary of thesis work Structure of thesis Contributions, self-refs, github links

# Chapter 2

# Word embeddings

## 2.1 Semantic encoding of words

## 2.2 State-of-the-art models for learning word embeddings

## 2.3 Multilingual word embeddings

### 2.3.1 Motivation

### 2.3.2 Tasks (e.g. MT)

### 2.3.3 Under-resourced languages

### 2.3.4 State-of-the-art models

**Smith**

**Facebook: MUSE**

## 2.4 Multilingual data

### 2.4.1 Panlex

### 2.4.2 Facebook embedding

# Chapter 3

# Proposed model

## 3.1 Description of our method

## 3.2 Software architecture

Used packages (tensorflow . . . )

# Chapter 4

# Experiments

## 4.1 Description of different tasks

### 4.1.1 Word translation

**Smith: eng-ita**

### 4.1.2 Cross-lingual semantic word similarity

**Facebook (MUSE): SemEval 2017**

### 4.2 Summarizing the contributions of the thesis

# Chapter 5

# Future work

# Köszönetnyilvánítás

Ez nem kötelezõ, akár törölhetõ is. Ha a szerzõ szükségét érzi, itt lehet köszönetet nyilvánítani azoknak, akik hozzájárultak munkájukkal ahhoz, hogy a hallgató a szakdolgozatban vagy diplomamunkában leírt feladatokat sikeresen elvégezze. A konzulensnek való köszönetnyilvánítás sem kötelezõ, a konzulensnek hivatalosan is dolga, hogy a hallgatót konzultálja.

# Bibliography

[1] Dan Jurafsky and James H Martin. *Speech and language processing*. Pearson London:, 2017.