MŰEGYETEM 1782

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 16, 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 16, 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 human-machine 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 Natural Language Processing [4], [1], [2]

Natural language processing is a vibrant interdisciplinary field with many different names, all reflecting a different facet of it. It is often referred to as speech and language processing, human language technology, computational linguistics, or speech recognition and synthesis. The goal of this field is to get computers to be able to understand and express themselves in human languages.

Natural Language Processing is hard since it deals with what is considered to be filled with many distinctive properties of the human being. This field is strongly connected with artificial intelligence since the way humans think and feel about the world is mainly happening in terms of human languages.

Being nowhere near as fast as digital communication, expressing ourselves with the help of human languages is a very effective way, though. Saying only the minimum message our listeners can fill up the rest with their world and common knowledge, and with the context of the situation. On the other hand, for a computer dealing with ambiguities and left out information is a very difficult issue to overcome.

The importance of computer integrated human language communication went as far as assigning truly intelligent machines to the ability of being capable of processing language as skillfully as we humans do. This idea was first considered by Alan Turing (1950) who introduced what has come to be known as the Turing test.

1.1.1 Common tasks

Natural language processing comprises a wide variety of task. Some of them like spam detection, POS-tagging, or named entity recognition are considered to be mostly solved problems. Applications for these tasks are now being in the market and are usually integrated to our smart devices even by default.

With some other tasks great progress has been made recently witch implies the existence of already fair enough applications but means that research work is still being done. Among them there are tasks like sentiment analysis, words sense disambiguation, syntactic parsing, machine translation etc.

What still really is considered to be very hard is to understand the meaning of a text. There are numerous interesting tasks for example question answering, dialogues, summarization, paraphrases, or text inference, for which in order to make relevant progress dealing with the semantics is inevitable.

1.1.2 Motivation for NLP research

We do really live in an era where natural language processing is gaining from day to day more and more attention. With the advent of smart phones the importance of language has gone even higher. These devices have small and rather inconvenient keyboards for which speech-driven communication seems very compelling. Big companies like Amazon, Apple, Facebook, Google, etc. are all putting out of products that use natural language to communicate with users. The contributions of this thesis are aiming the research field of word meaning and universal semantic representations, so below I will only mention applications that can directly take advantages of these contributions.

Speech-driven assistance applications can really make our everyday life more comfortable and more convenient. Let's just consider the fact how much help they could provide to people living with disabilities. These systems include speech input for which first automatic speech recognition technologies should be applied. But in order to understand the goal of the user we must run a semantic analysis.

An early version of conversational agents and certain strongly domain-based chatbots are already out on the market, providing 24 hour, immediate assistance for customers. Letting computers do the monotone and not at all creative tasks can leave to employees having more interesting jobs what only humans can do. or less working hours.

Advances of machine translation has led enjoy the English-based web services for non-English-speakers. For widespread languages this technology has already reached a well usable level, however, rare languages are still facing lack of coverage.

There are also numerous Web related tasks such as Web-based question answering which is basically an extended version of Web search where instead of key words we would be able to ask complete questions and communicate with the search engine just as like humans do among themselves. For all these applications programs should look way behind the syntactic surface and dig deep into the underlying semantics.

1.2 Universal Embeddings (1.5 page)

A commonly disputed psychological question, i.e. whether human conceptual structure is universal, is now being discussed not only by cognitive scientists ([6], [5], [3]), but more frequently explanations are being aided by computational methodologies as well. E.g. Youn et al. in [7] has shown that the human conceptual structure is independent of certain non-linguistic factors such as geography, climate, topology or literary traditions. Based on this knowledge we propose a procedure to find a universal semantic representation in a form of a universal embedding along with translation matrices that serve to map each language into a universal space.

Summary of thesis work Structure of thesis (chapter 1, 2... blablabla) Contributions, self-refs, github links

Chapter 2

Word embeddings

2.1 Semantic encoding of words

Within the field of natural language processing a more specific area concentrates on semantic representations which are being leveraged both by classical semantic tasks such as question answering or chatbots and by other NLP tasks which in the strict sense of the word are not considered semantic tasks such as machine translation or syntactic parsing. A crucial part of all semantic tasks is to have a proper word representation which is capable of encoding the meaning as well.

One way to build a semantic representation is to use a distributional model, or as it is more commonly known a word embedding. Embeddings are word vectors in a high dimensional space, and aim to have the property that the more similar two words are the bigger their cosine similarity shall be. These vectors are trained on huge corpora using deep neural networks and their similarity is considered by far the most common source of information for semantic similarity in state-of-the-art systems.

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] https://www.youtube.com/playlist?list=PL3FW7Lu3i5Jsnh1rnUwq$_TcylNr7EkRe6$.

[2] URL: https://www.economist.com/technology-quarterly/2017-05-01/language.

[3] Peter Gärdenfors. *Conceptual spaces: The geometry of thought.* MIT press, 2004.

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

[5] George Lakoff. *Women, fire, and dangerous things.* University of Chicago press, 2008.

[6] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

[7] Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771, 2016.