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 22, 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 22, 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 has become a basic requirement. In order to meet this requirement it is inevitable 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.

The obtained translation matrices will be evaluated on different tasks. (2 sentences about results)

Chapter 1

Introduction

1.1 Natural Language Processing [8], [1], [2]

Natural language processing (NLP) 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 main 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 one of the most delicate characteristics of human being: with human language. 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 channels, expressing ourselves by means of human languages is a very effective way of communication, though. Saying only the minimum message our listeners can fill up the rest with their world and common knowledge, and can easily figure out the missing or misunderstood parts from the context of the situation. This way they are also able to resolve ambiguities, homonyms etc. without even noticing it. Nonetheless, these tasks for a computer are not at all that trivial.

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

1.1.1 Common tasks of NLP

Natural language processing comprises a wide variety of tasks. 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 out 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 has to be done. Among them there are tasks like sentiment analysis, words sense disambiguation, syntactic

parsing, machine translation etc.

What is still 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 are living in an era when natural language processing is becoming more and more integrated into our every life. With the advent of smart phones the importance of language has gone even further. These devices are having small and rather inconvenient keyboards thus speech-driven communication seems very appealing. Big companies like Amazon, Apple, Facebook, Google, etc. are all putting out of products that use natural language to communicate with users. Since the contributions of this thesis are aiming the research field of word meaning and universal semantic representations, below I will only list 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 at first automatic speech recognition technologies should be applied. But after that, in order to understand the goal of the user, we must run a semantic analysis as well.

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. By letting computers do the monotone and not at all creative tasks employees could have more interesting jobs, jobs that only humans can do, or they could have less working hours in a week. Any of these would be a great progress for the society.

Advances in machine translation has already created a world where non-English speakers can also enjoy the benefits of the English-based web services. Generally, we can say that for widespread languages machine translation has already reached a fairly usable state, for rare languages, however, it is still facing difficulties.

There are also numerous Web related tasks that are strongly relying on the semantic analysis of the text. One promising task is for example the Web-based question answering task which is basically an extended version of the classical Web search, whereby instead of searching just for key words it would also be possible to ask complete questions and thus communicate with the search engine just like human beings do. For all these applications, however, it is inevitable to look way behind the syntactic surface and dig deep into the underlying semantics.

1.2 Universal Embeddings (1.5 page)

Given the need for robust representations for many languages, the question of whether human conceptual structure is universal has recently gained interest not only among cognitive scientists ([15], [11], [6]), but among computational linguists as well. Youn et al. [16]

has shown that human conceptual structure is independent of certain non-linguistic factors such as geography, climate, topology or literary traditions. Based on such findings we propose a procedure to construct a universal semantic representation in form of translation matrices that serve to map each language to the universal space. We use the pretrained fastText word embeddings [5], which are available for 294 languages.

TODO: summary of experiments and results.

Github links, self-references, contributions

The thesis is structured as follows:

- **Chapter 1** explains briefly the goals and motivations of the research field of natural language processing. It also summarizes the main contributions and the results of the accomplished work.
- **Chapter 2** discusses the state-of-the-art semantic word representations, i.e. the word embedding. It briefly describes the standard learning procedure for monolingual word-emeddings (word2vec) and it introduces the concept and resources for multilingual word-embeddings.
- **Chapter 3** describes the proposed model. It details the learning procedure and discusses the basic infrastructural and architectural features.
- **Chapter 4** presents all the experiments. It summarizes our results and compares them with the performance of other systems.
- **Chapter 5** is denoted to the description of the future work. This chapter suggests modifications, follow-ups for which we didn't have time to accomplish, or which are beyond the scope of this thesis work.

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. The idea is based upon the observation that synonyms or words with similar meanings tend to occur in similar contexts. For example in the following two sentences ‘‘`The cat is walking in the bedroom`’’ and ‘‘`A dog was running in a room`’’ words like ‘‘`dog`’’ and ‘‘`cat`’’ have exactly the same semantic and grammatical roles therefore we could easily imagine the two sentences in the following variations: ‘‘`The dog is walking in the bedroom`’’ and ‘‘`A cat was running in a room`’’ [4]. Based upon this intuition, what distributional models are aiming to do is to compute the meaning of a word from the distribution of words around it. [8]. The obtained meaning representations are usually high dimensional vectors, called as word embeddings, which refers to their characteristic feature that they model a world by embedding it into a vector space.

Embeddings are not only proved to be a better alternative of n-grams used for language modelling [4], but they are doing quite well on semantic tasks too. Mikolov et al. [12] has shown that the characteristics of word embeddings go way beyond the simple syntactic regularities. They showed that applying simple vector operations (e.g. vector addiction and subtraction) can often produce meaningful results. For example it was shown that *vector(”King”) - vector(”Man”) + vector(”Woman”)* results in a vector that is closest to the vector representation of the word *Queen* [13]. Moreover, these days’ state-of-the-art results on work similarity tasks are all held by word embeddings, where the similarity of two words are calculated as the normalized dot product of the corresponding word vectors. This measure is the so-called cosine similarity of word embeddings.

Another way to build semantic representations is to utilize lexical databases. In our

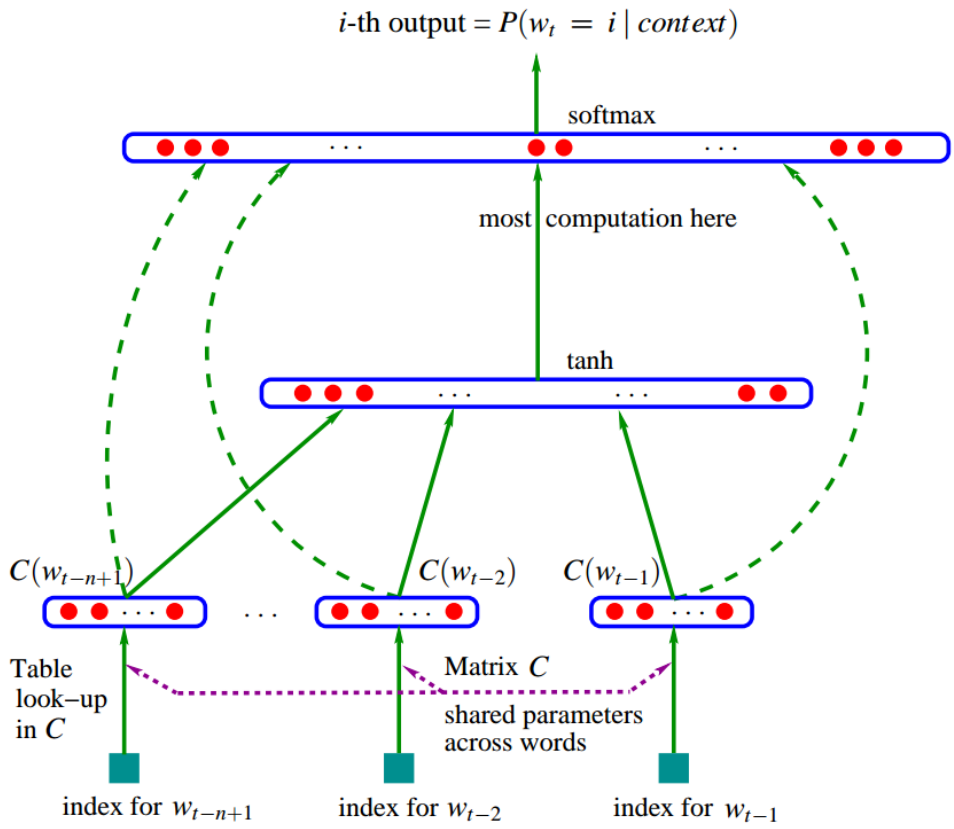
previous work we created a hybrid system leveraging both the `4lang` othologycal model described in [9], [10] and [3] and various distributional models, i.e. various word embeddings. With this system in 2016 we reached a state-of-the-art score on the `SimLex-999` [7] benchmark data [14].

In the following sections I will describe the basic procedure of training word embeddings and following that I will focus on a more specific field of computational semantics, on multilingual word embeddings.

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

In 2003 Bengio et al.[4] suggested a probabilistic feedforward neural network language model (NNLM) for learning a distributed representation for words. The network consists of input, projection, hidden and output layers, where at the input layer the N previous words are encoded using 1-of-V coding, where V is the size of the vocabulary. Figure 2.1 shows an overview of the architecture. This procedure was evaluated in terms of perplexity at which in comparison with the best of the n-grams it proved to be much better. The drawback of this architecture is that it becomes complex for computation between the projection and the hidden layer, as values in the projection layer are dense.

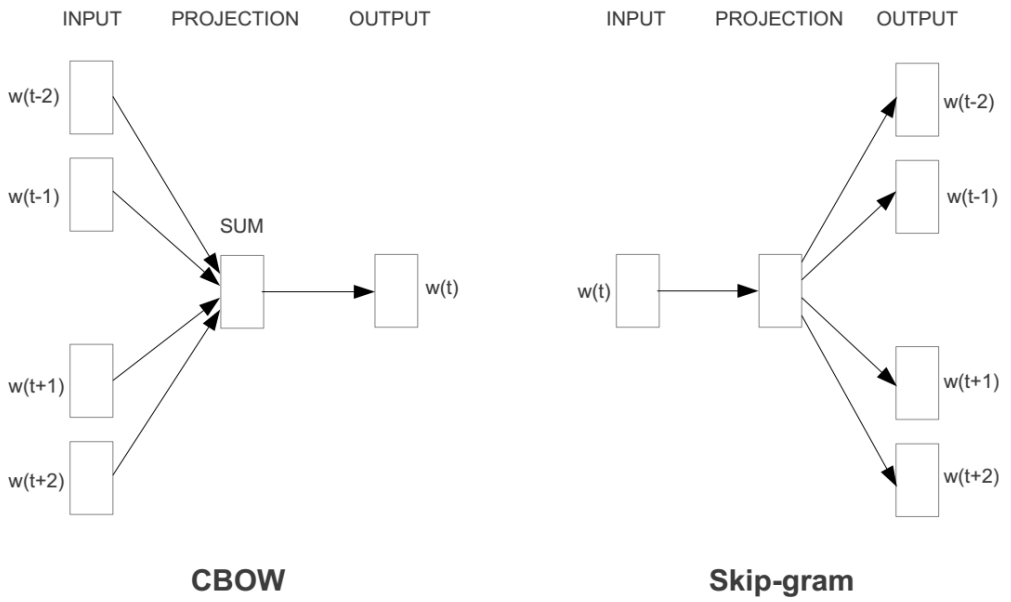
Figure 2.1: *Network architecture proposed by Bengio et al.[4]*



10 years later, in 2013 Mikolov suggested a bag-of-words neural networks, more specifically two different architecture. The first one, denoted as the CBOW (Continuous Bag-of-

Words Model) tries to predict the current word based on the context, whereas the second one, denoted as the continuous skip-gram model tries to maximize the classification of a word based on another word in the same sentence. Both models worked better than the NNLM suggested by Bengio [4] both on semantic and syntactic tasks, while between the two models of Mikolov the CBOW turned out to be slightly better on syntactic tasks and the skip-gram on semantic tasks. Mikolov procedure has become known as the `word2vec` procedure and it is available on github http://deeplearning4j.org/word2vec. The architecture of the CBOW and the skip-gram models are shown in figure 2.2.

Figure 2.2: *Bag-of-words neural networks suggested by Mikolov et al.[12]*



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$_T cylNr7EkRe$6.

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

[3] Judit Acs, Katalin Pajkossy, and András Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, 2013.

[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

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

[7] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

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

[9] András Kornai. The algebra of lexical semantics. In *the Mathematics of Language*, pages 174–199. Springer, 2010.

[10] András Kornai. Eliminating ditransitives. In *Formal Grammar*, pages 243–261. Springer, 2012.

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

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[13] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, 2013.

[14] Gábor Recski, Eszter Iklódi, Katalin Pajkossy, and Andras Kornai. Measuring semantic similarity of words using concept networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 193–200, 2016.

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

[16] 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.