Evaluating Bilingual Embeddings in Bilingual Dictionary Alignment

Çift Dilli Kelime Temsilleri ile Sözlük Eşlenmesi

Yiğit Sever

Dr. Gönenç Ercan

Supervisor

Submitted to
Graduate School of Science and Engineering of Hacettepe University
as a Partial Fulfilment to the Requirements
for the Award of the Degree of Master of Science
in Computer Engineering

2019

Declaration of Authorship

I, Yiğit Sever, declare that this thesis titled, “Evaluating Bilingual Embeddings in Bilingual Dictionary Alignment” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

Yiğit Sever

Evaluating Bilingual Embeddings in Bilingual Dictionary Alignment

Dictionaries catalog and describe the semantic information of a lexicon. WordNet provides an edge by presenting distinct concepts with the hierarchy information among them. Research in computer science has been using this hand crafted tool in natural language applications such as text summarization and machine translation. Original WordNet has been compiled for English yet counterparts for other languages are not as readily available nor as comprehensive. In order for research on languages other than English to benefit from the power of a WordNet, machine assisted creation and evaluation methods are essential.

Word embeddings can provide a mapping between words and points in a real valued vector space. Using these vectors, representing documents as well as forming geometric relationships between them is a well studied area of research. In this thesis we start by hypothesizing that a dictionary definition captures the semantic basis of the described word. We used word embeddings as building blocks to map dictionary definitions into a multidimensional space. These spaces can be aligned to accommodate two languages, allowing the transfer of information from one language to another. We investigate the success of retrieving and matching discrete senses across languages by employing supervised and unsupervised methods. Our experiments show that dictionary alignment can be evaluated successfully by using both unsupervised and supervised methods but corpora sizes should be taken into consideration. We further argue that some methods are not viable considering their poor performance.

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Contents

List of Figures

List of Tables

For/Dedicated to/To my...

Chapter 1

Introduction

1.1 Dictionaries

Dictionaries are living records of a society's language usage. Languages change over time, people adopt new words for new senses while others fall out of use. Concepts appear as a result of technological advancements or social shifts, giving birth to new senses and words to define them. Meanwhile, the term *dictionary* is a broad one to define. On its own, it brings forth the monolingual dictionary into consideration [1]. This type of dictionary presents words alongside their definitions following an alphabetical order. The intention is to inform the user about the words [2]. Other types of dictionaries vary with regard to their use case, target audience, and scope. For instance, bilingual dictionaries present words alongside their translations in the target language, often used by language learners or translators. Domain specific dictionaries list technical terms that target people who are familiar with the terminology.

The term that precedes the entries is called *headword* or *lemma*. Usually, lemmas are the form of a word without inflections. The sense they convey is as comprehensive as possible, reducing the number of otherwise redundant entries that would have been the derivatives of the unmarked form [3].

Dictionaries also inform the user about how senses relate to each other. **Polysemous** words share the same spelling while having related, often derivative meanings. For example; under the entry for the term *bank*, a definition might clarify the meaning *financial institution* while another can define *the building* *of a financial institution*. In contrast, **homonymous** words have distinct meanings while having identical spellings through coincidence. Formal definition of homonymy separates sound based and spelling based homonymy differently as homophones and homographs but for the purposes of our text based arguments, we do not delve into the specifics. The *bank of a river* is homonym to the given examples. Homonyms are often shown in discrete blocks of descriptions.

Synonymity is another lexical relation we are interested in. A word is synonymous to another if they share the same meaning but are not spelled alike, such as the terms *right* and *correct*. However, synonymity is seldom shown in dictionaries.

Dictionaries take an immense amount of time and expertise to prepare. We can talk about the examples after narrowing our scope down to the dictionaries that are still available today. A survey by Uzun [4] notes that the first instalment of the modern Turkish dictionary, led by a team of experts, has taken over 6 years to prepare. Kendall [5] talks about how Noah Webster, the writer of the *An American Dictionary of the English Language* had to mortgage off his home in order to finish his project which took over 26 years. The bulk of this effort is collecting documents and other written material in order to establish a *corpus* [4]. This endeavour is necessary since a corpus is crucial to create the vocabulary of a language. Once the corpus is at hand, researchers can extract the lemmas. The resulting wordstock is called the *lexicon* of the language.

The internet radically changed the way researchers aggregate data. The advancements in digital storage technology allowed the data to be persistent. Improvements in networking ensured that people can share the volume of it among themselves. With the popularization of social media, the internet generates everyday conversations at an unprecedented rate that researchers are using for natural language applications. Moreover, efforts on open, collaborative, web based encyclopedias generate structured, multilingual data often used in machine translation and text categorization tasks. Once the cumbersome task of corpus attainment is now akin to web crawling. With the digitized data, it was only natural for dictionaries to go digital as well since it's generally acknowledged that they are no longer viable if they are not electronic [1].

1.2 WordNet

George A. Miller started the WordNet project in the mid-1980s. On its early days, project members studied theories that were aimed towards enabling computers to understand natural language as intrinsically as humans do. While working on then popular semantic networks and sense graphs, they have started something that will evolve into an expansive, influential resource [6].

Traditional dictionaries are rigid, constrained by the nature of the printed form. Today, people can browse WordNet via queries, like an online dictionary or a thesaurus. Behind the scenes, a sprawling lexical database has relationship information for more than 117000 senses. Figure 1.1 shows a brief result for the query string "run".

Noun

S: (n) **run**, tally (a score in baseball made by a runner touching all four bases safely) *"the Yankees scored 3 runs in the bottom of the 9th"; "their first tally came in the 3rd inning"*

direct hyponym / full hyponym

- S: (n) earned run (a run that was not scored as the result of an error by the other team)
- S: (n) unearned run (a run that was scored as a result of an error by the other team)
- S: (n) run batted in, rbi (a run that is the result of the batter's performance) *"he had more than 100 rbi last season"*

direct hypernym / inherited hypernym / sister term

- S: (n) score (the act of scoring in a game or sport) *"the winning score came with less than a minute left to play"*

derivationally related form

- W: (v) run [Related to: run] (make without a miss)
- W: (v) tally [Related to: tally] (keep score, as in games)
- W: (v) tally [Related to: tally] (gain points in a game) *"The home team scored many times"; "He hit a home run"; "He hit .300 in the past season"*

S: (n) test, trial, **run** (the act of testing something) *"in the experimental trials the amount of carbon was measured separately"; "he called each flip of the coin a new trial"*

S: (n) footrace, foot race, **run** (a race run on foot) *"she broke the record for the half-mile run"*

Verb

- S: (v) **run** (move fast by using one's feet, with one foot off the ground at any given time) *"Don't run--you'll be out of breath"; "The children ran to the store"*
- S: (v) scat, **run**, scarper, turn tail, lam, run away, hightail it, bunk, head for the hills, take to the woods, escape, fly the coop, break away (flee; take to one's heels; cut and run) *"If you see this man, run!"; "The burglars escaped before the police showed up"*

Figure 1.1: WordNet result for the query "run", truncated for brevity.

WordNet lists terms, much like a traditional dictionary, alongside its polysemes but also their homonyms. Additionally, there is a horizontal association; for any sense, the lemmas that share the row with the target term are synonyms. This set of synonyms is aptly named *synsets*. A short description is also provided to clarify the meaning. These descriptions, hence the meanings for any synset is unique within the WordNet. During this discussion, we have used sense and synset interchangeably.

WordNet also includes other relationships such as *hypernymy* and *hyponymy*, semantic relation of senses being type-of one another [7].[1] For instance, the term "building" is a hyponym of "restaurant" since it encompasses a more general sense; the restaurant is type of a building. While coffee shop is a hypernym to the restaurant since it is a more specific sense. One other relation is the meronymy, defined as a sense being part of or a member of another [8]. Keeping to our building example, windows are meronym to buildings. Other relationships exist but listing them is outside the scope of this thesis. Bottom line is the effort that has gone through to map $117,000$ senses according to different semantic relationships. Sagot & Fišer [9] argue that the semantic relationships between senses are not tied to a specific language. With this assumption at hand, we can infer the effort behind the WordNet does not need to be repeated but can be translated to other languages.

Since it's inception, other projects built lexical databases, using the same WordNet design. Fellbaum [10] talks about the correct terminology that we abide for the thesis; "As WordNet became synonymous with a particular kind of lexicon design, the proper name shed its capital letters and became a common designator for semantic networks of natural languages". Hence *WordNet* refers to English Princeton WordNet, while *wordnets* created for other languages are not stylized.

[1] not to be confused with homonymy

1.3 Multilingual Wordnets

Authorities list more than 7000[2] living languages but only 40[3] of them have a sizeable presence on the internet. Among this small fraction, English is the dominant language of the web. English in not the centrepiece for natural language processing research because of any linguistic attribute. It is simply the most abundant language on web, giving researchers data to work with.

Natural language processing library spaCy [4] resorts to lemmatizations such as =PRON= to denote pronouns in order to collapse the senses for "I" "you", "them" etc.. The sense and the accompanying word for being the brother of a person's father or mother differs in Turkish while both collapse in "uncle" in English. Studying other languages can provide insight towards concepts that are not present in English.

Translation, information transfer from foreign languages is a valid way of enriching a language's corpora; if a term that for a sense does not have a match in the target language, it is a good indication for the linguists of that language to look into their lexicons and work towards expanding it [3]. Further research in the area contributes to languages other than English having access to tools that will incorporate them into the literature.

Open Multilingual WordNet [11] set out to discover the effects related to the choice of license for wordnets. Their criteria for usefulness is the number of citations a publication tied to the wordnet has gotten on literature. They identified two major problems with the current distributions;

- some projects have picked restrictive licenses, effectively barring access to their tools for research purposes.
- the structures of the wordnets are not standardized, creating additional cost for creating programs to parse and use the wordnets.

In order to overcome the standardization issue, Bond & Paik have aligned the wordnets according to their English Princeton WordNet lemma ids and have

[2] https://www.ethnologue.com/statistics
[3] https://w3techs.com/technologies/history_overview/content_language/
[4] https://spacy.io/

written individual scripts to parse them. They are currently hosting the results from a single source.[5]

With alignment information at hand, we have created our dataset that we will assume to be perfectly aligned; a golden corpus. Among the 34 wordnets available on Open Multilingual WordNet, only 6 of them have gloss information available. Given this thesis will only investigate the ability to map senses using definitions of the sense, we used the subset of Albanian [12], Bulgarian [13], Greek [14], Italian [15], Slovenian [16] and Romanian [17] wordnets. Table 1.1 shows brief statistics about them. We should note that the languages of the wordnets used in the thesis are all present in the 40 languages that have a significant presence on the internet that we have mentioned before. We have constrained this study to use only the freely available wordnets and not considered wordnets that are gated behind restrictive licenses.

Name of the Project	Language	Number of Definitions
Albanet	Albanian	4681
BulTreeBank WordNet	Bulgarian	4959
Greek Wordnet	Greek	18136
ItalWordnet	Italian	12688
Romanian Wordnet	Romanian	58754
SloWNet	Slovenian	3144

Table 1.1: Summary of the Wordnets used.

1.4 Thesis Goals

In this thesis, we will study document matching and document retrieval methods.

We will evaluate existing methods for their performance on cross-lingual document retrieval but our documents are dictionary definitions which are short, descriptive snippets of text. At the end of this study, we will answer the following research questions;

1. Is it possible to create wordnet like lexical databases using unsupervised document matching and retrieval techniques.

[5] http://compling.hss.ntu.edu.sg/omw/

2. How well does the studied techniques perform.
3. What attributes need to be considered regarding the available data.

1.5 Thesis Outline

`Fill later…`

Chapter 2

Background Information & Related Work

James Somers puts down the modern dictionaries in their article *You're Probably Using the Wrong Dictionary*. "The definitions are these desiccated little husks of technocratic meaningese, as if a word were no more than its coordinates in semantic space." [18]. Even though the author criticises the efficiency of the dictionary definitions, we will build a thesis on the idea that we can represent senses using their dictionary definitions.

2.1 Word Embeddings

Word embeddings feature vectors for words. Language modelling algorithms learn word embeddings and map a lexicon to show words or phrases as points on a multidimensional latent space. This representation allows researchers access to the tools of the broad literature in linear algebra and machine learning. Vectors are intuitive for humans to interpret, more so for machines. They can be compared or operated on algebraically to build more expansive representations.

The embeddings can be saved to the disk in matrix notation. Each row is labelled with a word or phrase that is denoted by the following n real numbers. Researchers have been sharing their models on the internet for other researchers to simply download and use them on their own applications. Word embeddings acquired this way are often called *pre-trained*. Some examples to the pre-trained word embeddings that are online as of writing this thesis are *word2vec*, *GloVe*, *Numberbatch* and *fastText*.

In this section, we will briefly present the history of word embeddings. Word embeddings is a sprawling subject that researchers has been building upon the ideas from probabilistic, statistical and neural network models. We have omitted approaches that are not used for our study and constrained ourselves only to the literature that lead up to the model we will use. At the end of the section, we will talk about the reasoning behind our decision to choose the model that we primarily use on our experiments; *fastText* [19].

2.1.1 History of Word Representations

In order to talk about how words can be mapped to a multidimensional space, we should first talk about how the idea that *they can* has came about.

Linguistic Background

In his 1954 article, Harris [20] introduced his ideas which later came to known as *distributional hypothesis* in the field of linguistics. He argued that similar words appear within similar contexts. The famous quote by Firth [21] captures the idea as; "You shall know a word by the company it keeps!" For instance, the semantic similarity between the terms *jacket* and *coat* can be theoretically shown since they will be accompanied by similar verbs, such as *wear*, *dry clean* or *hang*, and similar adjectives such as *warm* or *leather*. However, for a researcher to extract these rules by hand would be infeasible.

Even though Harris argued that "language is not merely a bag of words", using unordered collection of word counts to capture the semantic information will be used in the literature and be known as the *bag-of-words* hypothesis.

Vector Space Model

The history of word embeddings is tightly coupled with *vector space models*. Initially, vector space models appeared in the field of information retrieval. The intent was to extract vectors that represented *documents*. First vector space model developed by Salton *et al.* [22] was presented in "A Vector Space Model for Automatic Indexing". It was the first application of bag-of-words hypothesis on a corpus to extract semantic information [23]. Salton *et al.* presented the novel idea of a *document space*, consisting of fixed sized vectors as the columns of a term document matrix. The dimensions of the vectors were the whole vocabulary of the corpus.

In this space, a document D_i is represented using t distinct terms as a row vector;

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{it})$$

The weights for the index terms are calculated by using the *tf-idf* measure introduced by Jones [24]. *tf-idf* is the multiplication of two metrics;

tf the number of times a term k occurs in a document

idf the inverse of the number of documents that contain k.

Salton *et al.* presented their particular weighting scheme where the term frequency is multiplied by the following inverse document frequency for the term k.

$$IDF_k = \lceil \log_2 n \rceil - \lceil \log_2 d_k \rceil + 1$$

Where n is the number of documents in the collection and d_k is the number of documents that consists the term k. The weighting scheme was selected to "assign the largest weight to those terms which arise with high frequency in individual documents, but are at the same time relatively rare in the collection as a whole". They have cast their similarity function between documents as the inner product between their vectors which corresponds to the cosine similarity.

The vector space model allowed Salton *et al.* to handle the similarity between documents as the angle between two vectors. Above all, they have shown that there is merit to handling documents as real valued vectors.

2.1.2 Latent Semantic Analysis

Deerwester *et al.* [25] introduced latent semantic analysis in order to address a crucial problem with the vector space model. They have identified that the naive term document matrix approach cannot handle synonyms and homonyms due to the fact that vector space model requires the words to match exactly between the two documents. Synonymity is an issue because the query can have terms that have the same meaning as the target word, without getting matched. On the other hand, homonyms can match with an unrelated word. In order to address these issues, their model seeks the higher order latent semantic structure in order to learn the *similarity between words*.

Latent semantic analysis starts with a word co-occurrence matrix X. The terms of the matrix is weighted by some weighting scheme. While original study used raw term frequencies, *tf-idf* is a possibility while Levy *et al.* [26] reports point-wise mutual information (PMI) [27] as a popular choice. A term document matrix X is then factorized into three matrices using singular value decomposition [28];

$$X = T_0 S_0 D_0'$$

Where the columns of T_0 and D_0' are orthogonal to each other and S_0 is the diagonal matrix of singular values. The singular values of S_0 can be ordered by size to keep only the k largest elements, setting others to zero [25].

They talk about the use for the resulting matrix as follows;

> Each term or document is then characterized by a vector of weights indicating its strength of association with each of these underlying concepts. That is, the "meaning" of a particular term, query, or document can be expressed by k factor values, or equivalently, by the location of a vector in the k-space defined by the factors.

They have used this technique to solve document similarity task and briefly mentioned *word similarity*. Using latent semantic analysis to measure word similarity is later studied by Landauer & Dumais [29]

2.1.3 Building Upon Distributional Hypothesis

While Deerwester *et al.* studied relatedness between words using vectors, their approach used the whole document for the co-occurrence information while the focus was clearly still on the document similarity.

Schütze [30] proposed "to represent the semantics of words and contexts in a text as vectors" and built upon word co-occurrence. They theorized a context window of 1000 *characters* in order to consider words that are close to the target word instead of the whole document in co-occurrence calculations. However, they claimed that the computation power available was not suitable yet to fully tackle the task.

Lund & Burgess [31] took the challenge and experimented upon 160 million *words* taken from the Usenet, a precursor to the internet. They used a context size of 10 words and provided a method to obtain feature vectors to represent the meaning of words. However, intricate tuning of word co-occurrence generated associatively similar vectors instead of semantically similar ones.

2.1.4 Distributed Vector Representations

Bengio *et al.* [32] proposed learning word representations using a feedforward neural network. Their model would learn feature vectors for words using a predictive approach instead of counting based approaches we have presented until now. Although neural networks have been proposed to learn a language model by Xu & Rudnicky [33], the main contribution of Bengio *et al.* is to use an *embedding layer*, in order to attack *curse of dimensionality*. For a corpus of V words, there are $|V|$ dimensions for the language model to learn and taking *n-gram* representations into consideration, the problem grows exponentially. Using m dimensions in the embedding layer allowed Bengio *et al.* to represent words using manageable, more representative dimensions.

The setup for the neural network starts with the one hot encoded vector representation of the context for a word w. This context window is similar to those used in statistical models that predicts the word w_t using the words that lead up to w_t. In other words, context window of w_t is T words on the left of w_t.

$$\hat{P}(w_1^T) = \prod_{t=1}^{T} \hat{P}(w_t|w_1^{t-1})$$

The input layer is projected into an embedding layer, later to a softmax layer to get a probability distribution.

$$\hat{P}(w_t|w_{t-1},\ldots,w_{t-n+1}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{2.1}$$

However, this formulation is too expensive computationally since all vocabulary needs to be considered for the sum in the denominator. Authors reported training times around 3 weeks using 3 to 5 context window sizes and vocabulary sizes around 17000.

Collobert & Weston [34] suggested a deep neural network model in order to learn feature vectors for various natural language processing tasks. Their proposed approach for language model is important for our case since it explicitly learned distributed word representations or simply *word embeddings*. They have introduced two key ideas;

- Instead of using a context window that used words left of the target word to estimate the probability of the target word, they have placed the context window *on* the target window, using n words for left and right of the target word.
- They introduced negative examples, where they randomly changed the middle word with a random one. This allowed them to use the ranking cost;

$$\sum_{s\in S}\sum_{w\in D}\max\big(0, 1 - f(s) + f(s^w)\big)$$

Collobert & Weston [34] claim that these additions allowed their system to learn the representation better rather than the probability.

2.1.5 Pre-trained Embeddings

P. Turian *et al.* [35] evaluated the performance of different word representations as word features you can include into an existing task. They summarize their contribution as follows;

> Word features can be learned in advance in an unsupervised, task-inspecific, and model-agnostic manner. These word features, once learned, are easily disseminated with other researchers, and easily integrated into existing supervised NLP systems. [...] With this contribution, word embeddings can now be used off-the-shelf as word features, with no tuning.

2.1.6 Popularization of Word Embeddings

word2vec package [36–38] popularized word embeddings. There are two aspects of the work done by Mikolov *et al.* that contributed to the fact;

- Their model captures the semantic and syntactic attributes of words and phrases on a large scale with good accuracy, trained on billions of words.
- They published their code as well as their pre-trained embeddings as an open source project[1].

The second point is self explanatory but in order to argue about the first one, we should report the algorithms behind word2vec.

The *skip-gram model* [36] model differs from the previous methods by predicting the surrounding words given the target word. In "Distributed Representations of Words and Phrases and Their Compositionality", it is defined as follows;

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{2.2}$$

The $w_1, w_2, \ldots, w_T$ are the context of the word w_t. We should note that, Levy *et al.* [26] has identified that this window size is *dynamic* in the open source implementation of word2vec, where the actual window size is sampled between 1 and T.

Levy *et al.* [26] compared the performance of count based and prediction based word representation models. Representation algorithms they considered are;

- Positive pointwise mutual information (PPMI) [27, 39]
- Singular Value Decomposition on PPMI Matrix (Latent Semantic Analysis) [25]
- Skip-Gram with Negative Sampling [37]
- Global Vectors for Word Representation [40]

They found out that choice of a particular algorithm played an insignificant role compared to choosing the right *hyperparameters*. They used this finding to

[1] https://code.google.com/archive/p/word2vec/

counter the results reported by Baroni *et al.* [41]. Baroni *et al.* claimed that predictive models outperformed count based models. On the other hand, Levy *et al.* noted that Baroni *et al.* used count based models without hyperparameter tuning, denying them from "tricks" developed in the word representation literature.

2.1.7 fastText

Armed with the fact that a good word representation model should tune their hyperparameters and should be trained on a large dataset, we set our sights on *fastText* [2]. On their website, authors define fastText as (a) "Library for efficient text classification and representation learning". The ideas behind it are presented in Mikolov *et al.* [19]. Overall, it builds upon word2vec [37] by adding position dependent features presented in Mnih & Kavukcuoglu [42] and character n-grams suggested on Bojanowski *et al.* [43].

The character n-grams or subword model is presented in "Enriching Word Vectors with Subword Information". Instead of using a context window and learning the representation of a target word given the context or the other way around given the skip-gram model, Bojanowski *et al.* learn representations for character n-grams; With the subword vectors z_g for every n-gram g at hand, authors take a dictionary of n-grams of size G and for a given word w, they denote $G_w \subset 1, \ldots, G$ as the n-grams of w.

$$s(w, c) = \sum_{g \in G_w} z_g^T v_c$$

They are currently hosting pre-trained word embeddings for 157 languages, built from *Common Crawl* and *Wikipedia* data on their website.

2.2 Bilingual Word Embeddings

Cross lingual embedding models optimize similar objectives. Only source of variation is due to the data used and the monolingual regularization objectives employed [44].

[2] https://fasttext.cc/

2.3 Document Retrieval

2.4 Approaches in Wordnet Generation

WordNet generation is broken down into 4 categories

1. Expand model, Vossen [45], fixed synsets are translated from English to target language.
2. Link English entries from machine-readable bilingual dictionaries to English Princeton WordNet senses Knight & Luk [46].
3. Taxonomy parsing Farreres *et al.* [47].
4. Ontology matching Farreres *et al.* [48]

Sagot & Fišer [9] built a French wordnet.

Gordeev *et al.* [49] uses unsupervised cross-lingual embeddings to match cross-lingual product classifications. Working on taxonomy matching, they use out of domain pre-trained embeddings due to small size of their corpora and investigate methods using untranslated and translated text.

Lesk [50] represent words using their gloss. Relied upon traditional dictionaries. Banerjee & Pedersen [51] developed on lesk algorithm and included WordNet definitions. Khodak *et al.* [52] used word embeddings and WordNet.

Metzler *et al.* [53] talked about short text retrieval and lexical matching. They reported that lexical matching is good for finding semantically identical matches.

Xiao & Guo [54] another embedding paper.

Kusner *et al.* [55] is Word Mover's Distance.

Balikas *et al.* [56] suggested using optimal transport for cross-lingual document retrieval.

Arora *et al.* [57] simple but tough-to-beat baseline for sentence embeddings.

Klementiev *et al.* [58] base paper for cross lingual word embeddings?

Chapter 3

Unsupervised Matching

3.1 Linear Assignment Using Sentence Embeddings

Using word embeddings to obtain embeddings for longer pieces of text has been studied with implementations like doc2vec [59] that builds upon the word2vec [37] model in order to learn paragraph embeddings. However, there is an assumption of a continuous text for the given model. When the text that we would like to show on a latent space is not part of a longer piece of text but *discrete* pieces, that presumption does not hold. With the dictionary definitions, we have such a case. Our dictionary definitions are comprised of 10 to 11 words and there is no relation from one distinct dictionary definition to another. In other words, they are not continuous. One other case where a similar situation occur is *twitter*. *Tweets* are short pieces of text due to the 280 character constraint imposed by the platform. With such short pieces of text, instead of paragraph embeddings, we can talk about *sentence embeddings*. A sentence embedding model should ideally capture the collective meaning of the short text where every word is potentially informative.

Zhao *et al.* [60] used two approaches for SemEval-2015 Task 2: Semantic Textual Similarity [1]. First, for a sentence $S = (w_1, w_2, \ldots, w_s)$ where the length of the presumably small sentence is $|S| = s$ and the word embedding of a w_t is v_t;

- They summed up the word embeddings of the sentence $\sum_{t \in S} v_t$
- Used information content [61] to weigh each word's LSA vector $\sum_{t \in S} I(w_t) v_t$

[1] http://alt.qcri.org/semeval2015/task2/

Table 3.1: Some definitions from English Princeton WordNet

turn red, as if in embarrassment or shame
a feeling of extreme joy
a person who charms others (usually by personal attractiveness)
so as to appear worn and threadbare or dilapidated
a large indefinite number
distributed in portions (often equal) on the basis of a plan or purpose
a lengthy rebuke

Both approaches results in a vector that is in the same dimensions R^d as the original word representations.

Edilson A. Corrêa *et al.* [62] expanded upon this simple yet effective idea to tackle the SemEval-2017 Task 4[2], Sentiment Analysis in Twitter. In order to acquire embeddings that represented *tweets*, they weighed the word embeddings that made up a tweet; tweet$_i = (w_{i1}, w_{i2}, \ldots, w_{im})$ with the *tf-idf* weights. For the *tf-idf* calculation, they cast individual weights as documents so that term frequency become the term count in a single tweet while document frequency become the number of tweets the term w_t occurs.

We have mentioned that our dictionary definitions are not continuous. Yet, we advocate using *tf-idf* weights to weigh our word embeddings to get sentence embeddings. In order to clarify, let us present Table 3.1.

turn red, as if in embarrassment or shame
a feeling of extreme joy
a person who charms others (usually by personal attractiveness)
so as to appear worn and threadbare or dilapidated
a large indefinite number
distributed in portions (often equal) on the basis of a plan or purpose
a lengthy rebuke

For the *tf-idf* calculations, we followed a similar approach. The term frequency is the raw count of a term in a dictionary definition. While the document frequency is the number of dictionary definitions where w_t occurs.

[2] http://alt.qcri.org/semeval2017/task4

Then, with the term-embedding matrix at hand, we have calculated definition embeddings using;

$$S_{\text{emb}}(S) = \sum_{w_i \in S} \mathit{tf}_{w_i,S}\text{-}\mathit{idf}_{w_i} \cdot Emb_w(w_i) \tag{3.1}$$

Every word that makes up a definition is scaled by its vector in $\mathbb{R}^n$, then concatenated to form sentence embeddings on $\mathbb{R}^n$.

Given the N vectors from source and target language, we hypothesize that there exists a matching where every source definition vector is perfectly mapped to one target vector. Given that this problem naively iterates over $N!$ matchings, we have looked into an algorithm.

3.2 Jonker Volgenant Algorithm

3.3 Results

	Percentage of Correctly Matched Definitions		
Language	**fastText 1M**	**fastText 500k**	**Numberbatch**
bg	0.39	0.41	0.19
el	0.37	0.38	0.14
it	0.28	0.28	0.36
ro	0.39	0.39	0.20
sl	0.15	0.15	0.06
sq	0.55	0.54	0.27

Table 3.2: Linear Assignment Using 2000 Definitions

	Percentage of Correctly Matched Definitions		
Language	**fastText 1M**	**fastText 500k**	**Numberbatch**
bg	0.35	0.36	0.18
el	0.36	0.36	0.12
it	0.25	0.25	0.32
ro	0.36	0.37	0.19
sl	0.11	0.11	0.05
sq	0.39	0.40	0.19

Table 3.3: Linear Assignment Using 3000 Definitions

	fastText 1M	**fastText 500k**	**Numberbatch**
Best	0.55	0.54	0.36
Worst	0.11	0.11	0.05
Average	0.33	0.33	0.19

Table 3.4: Summary of Linear Assignment

Chapter 4

Dictionary Alignment as Pseudo-Document Retrieval

Document retrieval is the prototypical information retrieval task. Bush [63] first theorized the possibilities of the automatic information retrieval by machines in his essay titled "As We May Think". Singhal [64] also gives due credit to Luhn [65] for the suggestion of document retrieval using word overlap.

Modern information retrieval techniques are far from the scope of this thesis. Considering the small collection of documents at hand, we will investigate if we can handle the task using approaches that were available to the researchers when the size of corpora that were available to them was small as well [64]. However, we will get leverage from a state of the art tool from the modern computer science that is Google Translate.

4.1 Machine Translation

The first method we will study starts off by translating the target language's corpora to English using Google Translate. We have used the Google Cloud API[1] in order to automate the process.

With the English Princeton WordNet definitions and the target wordnet definitions at hand, we can handle the task as monolingual document retrieval. We have used the vector space representation we have talked about in Chapter 2.

[1] https://cloud.google.com/translate

We have chosen *tf-idf* as to ask if the task at hand can be solved by naive tools. In order to get *tf-idf* scores of the documents, first a term-document matrix is created. Documents being definitions and with an average of 10.62 words per definition, the resulting matrix is parse. In a *tf-idf* matrix, for an entry in the matrix $w_{i,j}$, we can give the formula for it as:

$$\mathit{tf}_{w,d}\text{-}\mathit{idf}_w = \sum_{w' \in d} f_{w',d} \cdot \log \frac{N}{\mathit{df}_w}$$

Such that term $w_{i,j}$ depicts the importance of term t with relation to its general importance throughout the corpus. Now we can define the similarity between the documents as the cosine similarity between their *tf-idf* vectors. For the row w_t and w_p, cosine similarity between definitions t, p is

$$\cos(\theta) =$$

Definitions are then separated into queries and corpora. Query definitions is then matched up against every definition in the corpora and the ten documents that are closest in terms of cosine similarity is retrieved. Within the retrieved documents, if the document with the matched sense id is retrieved in the first result, this is taken as a hit at 1. Mean Reciprocal Rank is also calculated in order to show the success of a retrieval scenario.

Where monolingual retrieval falls short, we leveraged the power of word embeddings to capture the semantic information of the words. A famous example for the inadequacy of *tf-idf* is illustrated by [55]. For two snippets of text; *Obama speaks to the media in Illinois* and *The President greets the press in Chicago* Kusner argues that while they convey the same information, they would be near orthogonal in a bag of words setting. Yet before moving forward with WMD, we wanted to test sentence embeddings.

4.2 Cross Lingual Document Retrival

4.2.1 Optimal Transport

4.2.2 Sinkhorn

Chapter 5

Supervised Validation

5.1 Main Section 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

5.1.1 Subsection 1

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

5.1.2 Subsection 2

Morbi rutrum odio eget arcu adipiscing sodales. Aenean et purus a est pulvinar pellentesque. Cras in elit neque, quis varius elit. Phasellus fringilla, nibh eu tempus venenatis, dolor elit posuere quam, quis adipiscing urna leo nec orci. Sed nec nulla auctor odio aliquet consequat. Ut nec nulla in ante ullamcorper aliquam at sed dolor. Phasellus fermentum magna in augue gravida cursus. Cras sed pretium lorem. Pellentesque eget ornare odio. Proin accumsan, massa viverra cursus pharetra, ipsum nisi lobortis velit, a malesuada dolor lorem eu neque.

5.2 Main Section 2

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

Chapter 6

Experiments and Evaluation

We experimented with in domain *fasttext* embeddings. We have trained Romanian and Bulgarian embeddings and ran the WMD and Sinkhorn experiments. The performance dropped to the quarter of pre-trained embeddings so we have not repeated the experiments for other language corpora.

We have evaluated our embeddings using the standard bilingual lexicon extraction as a general measure for their performance. Although Ruder *et al.* [44] and Glavas *et al.* [66] both say you should evaluate them using downstream tasks.

Language	**FastText 1M**	**FastText 500k**	**numberbatch**
bg	33.61	35.17	51.97
el	37.37	39.58	30.35
it	58.20	59.28	50.37
ro	37.33	38.71	64.17
sl	21.42	22.91	74.74
sq	24.46	25.36	58.63

Table 6.1: Accuracy Scores of the Vectors Aligned Using VecMap

Language	FastText 1M	FastText 500k	numberbatch
bg	96.43	93.36	17.53
el	94.44	90.28	12.15
it	97.93	95.97	41.08
ro	97.06	94.91	16.4
sl	94.67	90.73	9.23
sq	83.59	80.92	9.51

Table 6.2: Coverage Scores for the Vectors Aligned Using VecMap

Chapter 7

Conclusion

Appendix A

Frequently Asked Questions

A.1 How do I change the colors of links?

The color of links can be changed to your liking using:

`\hypersetup{urlcolor=red}`, or

`\hypersetup{citecolor=green}`, or

`\hypersetup{allcolor=blue}`.

If you want to completely hide the links, you can use:

`\hypersetup{allcolors=.}`, or even better:

`\hypersetup{hidelinks}`.

If you want to have obvious links in the PDF but not the printed text, use:

`\hypersetup{colorlinks=false}`.

Bibliography

1. A *Practical Guide to Lexicography* (ed van Sterkenburg, P.) *Terminology and Lexicography Research and Practice* **6.** OCLC: 249659375 (Benjamins, Amsterdam, 2003). 459 pp. isbn: 978-90-272-2329-6 978-90-272-2330-2 978-1-58811-380-1 978-1-58811-381-8.
2. Uzun, E. N. Modern Dilbilim Bulguları Işığında Türkçe Sözlüğe Bir Bakış. *Çukurova Üniversitesi Türkoloji Araştırmaları Merkezi* (2005).
3. İbrahim USTA, H. Türkçe Sözlük Hazırlamada Yöntem Sorunları. *Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi*, 223–242 (Jan. 1, 2006).
4. Uzun, L. 1945'TEN BU YANA TÜRKÇE SÖZLÜKLER. KEBİKEÇ *İnsan Bilimleri İçin Kaynak Araştırmaları Dergisi*, 53–57. issn: 1300-2864 (1999).
5. Kendall, J. *The Forgotten Founding Father: Noah Webster's Obsession and the Creation of an American Culture* 1st Edition. 368 pp. isbn: 0-399-15699-2 (G.P. Putnam's Sons, Apr. 14, 2011).
6. Fellbaum, C. *WordNet : An Electronic Lexical Database* isbn: 978-0-262-27255-1 (MIT Press, 1998).
7. Miller, G. A. Nouns in WordNet: A Lexical Inheritance System. *Int J Lexicography* **3,** 245–264. issn: 0950-3846. https://academic.oup.com/ijl/article/3/4/245/923281 (2019) (Dec. 1, 1990).
8. Winston, M. E., Chaffin, R. & Herrmann, D. A Taxonomy of Part-Whole Relations. *Cognitive Science* **11,** 417–444. issn: 1551-6709. https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1104_2 (2019) (1987).
9. Sagot, B. & Fišer, D. Building a Free French Wordnet from Multilingual Resources (May 31, 2008).
10. Fellbaum, C. in *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* (ed Vossen, P.) 137–148 (Springer Netherlands, Dordrecht, 1998). isbn: 978-94-017-1491-4. https://doi.org/10.1007/978-94-017-1491-4_6 (2019).
11. Bond, F. & Paik, K. A Survey of WordNets and Their Licenses (Jan. 1, 2012).

12. Ruci, E. *On the Current State of Albanet and Related Applications* (Technical report, University of Vlora.(http://fjalnet. com ..., 2008).
13. Simov, K. I. & Osenova, P. *Constructing of an Ontology-Based Lexicon for Bulgarian.* in *LREC* (Citeseer, 2010).
14. Stamou, S., Nenadic, G. & Christodoulakis, D. *Exploring Balkanet Shared Ontology for Multilingual Conceptual Indexing.* in *LREC* (2004).
15. Pianta, E., Bentivogli, L. & Girardi, C. MultiWordNet: Developing an Aligned Multilingual Database (Jan. 1, 2002).
16. Fišer, D., Novak, J. & Erjavec, T. *sloWNet 3.0: Development, Extension and Cleaning* in *Proceedings of 6th International Global Wordnet Conference (GWC 2012)* (2012), 113–117.
17. Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A. & Ştefănescu, D. *Romanian Wordnet: Current State, New Applications and Prospects* in *Proceedings of 4th Global WordNet Conference, GWC* (2008), 441–452.
18. Somers, J. *You're Probably Using the Wrong Dictionary* `http://jsomers.net/blog/dictionary` (2019).
19. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. & Joulin, A. *Advances in Pre-Training Distributed Word Representations* in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (2018).
20. Harris, Z. S. Distributional Structure. *WORD* **10,** 146–162. issn: 0043-7956. `https://doi.org/10.1080/00437956.1954.11659520` (2019) (Aug. 1, 1954).
21. Firth, J. R. A Synopsis of Linguistic Theory 1930–1955. *Studies in linguistic analysis* **Special volume of the Phiological Society,** 11 (1957).
22. Salton, G., Wong, A. & Yang, C. S. A Vector Space Model for Automatic Indexing. *Commun. ACM* **18,** 613–620. issn: 0001-0782. `http://doi.acm.org/10.1145/361219.361220` (2019) (Nov. 1975).
23. Turney, P. D. & Pantel, P. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* **37,** 141–188. issn: 1076-9757. arXiv: `1003.1141`. `http://arxiv.org/abs/1003.1141` (2018) (Feb. 27, 2010).
24. Jones, K. S. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation* **28,** 11–21 (1972).

25. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. Indexing by Latent Semantic Analysis. *Journal of the American society for information science* **41,** 391–407 (1990).
26. Levy, O., Goldberg, Y. & Dagan, I. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* **3,** 211–225 (Dec. 1, 2015).
27. Church, K. W. & Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* **16,** 22–29. issn: 0891-2017. http://dl.acm.org/citation.cfm?id=89086.89095 (2019) (Mar. 1990).
28. Forsythe, G. E., Malcolm, M. A. & Moler, C. B. *Computer Methods for Mathematical Computations* (Prentice-hall Englewood Cliffs, NJ, 1977).
29. Landauer, T. K. & Dumais, S. T. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological review* **104,** 211 (1997).
30. Schütze, H. *Dimensions of Meaning* in *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing* Minneapolis, Minnesota, USA (IEEE Computer Society Press, Los Alamitos, CA, USA, 1992), 787–796. isbn: 978-0-8186-2630-2. http://dl.acm.org/citation.cfm?id=147877.148132 (2019).
31. Lund, K. & Burgess, C. Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence. *Behavior Research Methods, Instruments, & Computers* **28,** 203–208. issn: 1532-5970. https://doi.org/10.3758/BF03204766 (2019) (June 1, 1996).
32. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A Neural Probabilistic Language Model. *Journal of machine learning research* **3,** 1137–1155 (Feb 2003).
33. Xu, W. & Rudnicky, A. *Can Artificial Neural Networks Learn Language Models?* in *Sixth International Conference on Spoken Language Processing* (2000).
34. Collobert, R. & Weston, J. A *Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning* in. Proceedings of the 25th International Conference on Machine Learning (ACM, May 7, 2008), 160–167. isbn: 978-1-60558-205-4. http://dl.acm.org/citation.cfm?id=1390156.1390177 (2019).
35. P. Turian, J., Ratinov, L.-A. & Bengio, Y. *Word Representations: A Simple and General Method for Semi-Supervised Learning.* in. Proceedings of the

48th Annual Meeting of the Association for Computational Linguistics. **2010** (Jan. 1, 2010), 384–394.

36. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. arXiv: 1301.3781 [cs]. http://arxiv.org/abs/1301.3781 (Jan. 16, 2013).
37. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. arXiv: 1310.4546 [cs, stat]. http://arxiv.org/abs/1310.4546 (2019) (Oct. 16, 2013).
38. Mikolov, T., Yih, W.-t. & Zweig, G. *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* (2013), 746–751.
39. Bullinaria, J. A. & Levy, J. P. Extracting Semantic Representations from Word Co-Occurrence Statistics: A Computational Study. *Behavior Research Methods*, 510–526 (2007).
40. Pennington, J., Socher, R. & Manning, C. *Glove: Global Vectors for Word Representation* in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014), 1532–1543.
41. Baroni, M., Dinu, G. & Kruszewski, G. *Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors* in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* Baltimore, Maryland (Association for Computational Linguistics, June 2014), 238–247. https://www.aclweb.org/anthology/P14-1023.
42. Mnih, A. & Kavukcuoglu, K. in *Advances in Neural Information Processing Systems 26* (eds Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) 2265–2273 (Curran Associates, Inc., 2013). http://papers.nips.cc/paper/5165-learning-word-embeddings-efficiently-with-noise-contrastive-estimation.pdf.
43. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. https://arxiv.org/abs/1607.04606 (2018) (July 15, 2016).
44. Ruder, S., Vulić, I. & Søgaard, A. A Survey Of Cross-Lingual Word Embedding Models. arXiv: 1706.04902 [cs]. http://arxiv.org/abs/1706.04902 (2018) (June 15, 2017).

45. Vossen, P. Introduction to EuroWordNet. *Computers and the Humanities* **32,** 73–89. issn: 0010-4817. https://www.jstor.org/stable/30200456 (2019) (1998).
46. Knight, K. & Luk, S. K. *Building a Large-Scale Knowledge Base for Machine Translation* in *In Proceedings of AAAI* (1994).
47. Farreres, X., Rigau, G. & Rodríguez, H. *Using WordNet for Building WordNets* /paper/Using-WordNet-for-Building-WordNets-Farreres-Rigau/01405726f0dac56b063cb4ee04e766daf3993c23 (2019).
48. Farreres, J., Gibert, K. & Rodríguez, H. *Towards Binding Spanish Senses to Wordnet Senses* (2004).
49. Gordeev, D., Rey, A. & Shagarov, D. *Unsupervised Cross-Lingual Matching of Product Classifications* in *Proceedings of the 23rd Conference of Open Innovations Association FRUCT* Bologna, Italy (FRUCT Oy, 2018), 62:459–62:464. http://dl.acm.org/citation.cfm?id=3299905.3299967.
50. Lesk, M. *Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone* in *Proceedings of the 5th Annual International Conference on Systems Documentation* (ACM, New York, NY, USA, 1986), 24–26. isbn: 978-0-89791-224-2. http://doi.acm.org/10.1145/318723.318728.
51. Banerjee, S. & Pedersen, T. *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet* in *Computational Linguistics and Intelligent Text Processing* (ed Gelbukh, A.) (Springer Berlin Heidelberg, 2002), 136–145. isbn: 978-3-540-45715-2.
52. Khodak, M., Risteski, A., Fellbaum, C. & Arora, S. *Automated WordNet Construction Using Word Embeddings* in *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and Their Applications* (2017), 12–23.
53. Metzler, D., Dumais, S. & Meek, C. *Similarity Measures for Short Segments of Text* in *Advances in Information Retrieval* (eds Amati, G., Carpineto, C. & Romano, G.) (Springer Berlin Heidelberg, 2007), 16–27. isbn: 978-3-540-71496-5.
54. Xiao, M. & Guo, Y. *Distributed Word Representation Learning for Cross-Lingual Dependency Parsing* in (Jan. 1, 2014), 119–129.

55. Kusner, M. J., Sun, Y., Kolkin, N. I. & Weinberger, K. Q. *From Word Embeddings to Document Distances* in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37* Lille, France (JMLR.org, 2015), 957–966. http://dl.acm.org/citation.cfm?id=3045118.3045221 (2019).
56. Balikas, G., Laclau, C., Redko, I. & Amini, M.-R. Cross-Lingual Document Retrieval Using Regularized Wasserstein Distance. arXiv: 1805.04437 [cs, stat]. http://arxiv.org/abs/1805.04437 (2019) (May 11, 2018).
57. Arora, S., Liang, Y. & Ma, T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. https://openreview.net/forum?id=SyK00v5xx (2019) (Nov. 4, 2016).
58. Klementiev, A., Titov, I. & Bhattarai, B. *Inducing Crosslingual Distributed Representations of Words* (2012).
59. Le, Q. V. & Mikolov, T. Distributed Representations of Sentences and Documents. arXiv: 1405.4053 [cs]. http://arxiv.org/abs/1405.4053 (May 16, 2014).
60. Zhao, J., Lan, M. & Tian, J. ECNU: *Using Traditional Similarity Measurements and Word Embedding for Semantic Textual Similarity Estimation* in *SemEval@NAACL-HLT* (2015).
61. Šarić, F., Glavaš, G., Karan, M., Šnajder, J. & Bašić, B. D. *TakeLab: Systems for Measuring Semantic Text Similarity* in *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* Montréal, Canada (Association for Computational Linguistics, Stroudsburg, PA, USA, 2012), 441–448. http://dl.acm.org/citation.cfm?id=2387636.2387708 (2019).
62. Edilson A. Corrêa, J., Marinho, V. & Borges dos Santos, L. NILC-USP at SemEval-2017 Task 4: A Multi-View Ensemble for Twitter Sentiment Analysis (Apr. 7, 2017).
63. Bush, V. As We May Think. *The atlantic monthly* **176,** 101–108 (1945).
64. Singhal, A. Modern Information Retrieval: A Brief Overview. *Bulletin of the Ieee Computer Society Technical Committee on Data Engineering* **24,** 2001 (2001).

65. Luhn, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of research and development* **1,** 309–317 (1957).
66. Glavas, G., Litschko, R., Ruder, S. & Vulic, I. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. arXiv: 1902.00508 [cs]. http://arxiv.org/abs/1902.00508 (2019) (Feb. 1, 2019).