Proposal: Comparing Unsupervised Machine Translation Strategies with Word2Vec

II2202, Fall 2015

Bram Leenders
KTH, Royal Institute of Technology
Brinellvägen 8, 114 28 Stockholm
Sweden
b.c.leenders@gmail.com

Marc Romeyn
KTH, Royal Institute of Technology
Brinellvägen 8, 114 28 Stockholm
Sweden
marc.romeyn@gmail.com

Keywords
Word2Vec, Machine Translation

1. INTRODUCTION
The Internet offers a vast amount of natural language that can be used in natural language processing, for a very low price. A study by Buck et al. [2] estimated that each of the top 10 most frequently used languages on the internet has at least 250 GiB worth of text publically available online. For English (the most frequent), they even found 23 TiB of text.

Such amounts of text have a big potential to be used for the training of language models, but only if building these models can be done in an unsupervised fashion. Manually curating, marking and tagging text is far too much work to be feasible. With unsupervised algorithms, however, the structure in language can be exploited to let computers build language models.

1.1 Problem Statement
This project will focus on unsupervised training of computer models to translate words. Specifically, we focus on the use of word2vec [5] to provide translations.

Given the large bodies of text publically available, we want to provide automated translations.

1.2 Prior Work
Since the introduction of word2vec [5, 7] in 2013, the algorithm has seen a wide variety of usecases. In the initial paper [5], Mikolov et. al describe interesting relations between vectors corresponding to words. A famous example of how word2vec models relations between words as mathematical equations is $king - man + woman = queen$. The sematic relationships between man/woman and king/queen are preserved in the transformation of words to vectors, and can be expressed with basic algebra.

Subsequent papers have improved the algorithm both in terms of accuracy ([4]), performance, parallelization and extended the initial scope of applications. A good example of the latter is a paper by Boycheva [1], which uses word2vec outside the natural language processing (NLP) domain but to generate playlists. Based on a set of playlists, their word2vec-based algorithm can suggest new playlists with artists that go well together.

One of the applications of word2vec inside the NLP domain, is exploiting similarities in languages for assistance in machine translation [8]. Mikolov et al. [6] released a subsequent paper on word2vec, in which they describe similarities between models of different languages. An example they give, is how the usage of the numbers one to five in English is very similar to the usage in Spanish, and likewise for the names of animals. Figure 1 shows a graphical representation of word vectors in English and Spanish.

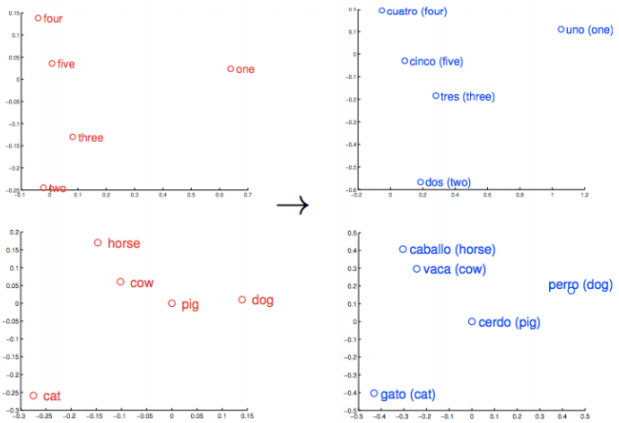

Figure 1: Vector representations of English and Spanish words, after dimensionality reduction and rotation. Notice the high level of similarity between both languages. Reprinted from Mikolov et al. [6]

The similarities between languages can be used to predict translations for words without any human interaction or labeled input data. Using only unsupervised machine learning techniques, a computer could learn how to translate English to for instance Spanish and vice versa. The only requirement is a large amount of text in both languages to train the word2vec models on.

In this research, we will focus on this specific application of word2vec: using similarities in languages to provide translations of words.

It is important to note that word2vec only uses information of co-occurrences to model words. It does not learn grammatical concepts other than by statistical analysis. This limits our translation to single words; although the translator might be able to translate each word individually, it cannot learn that each finite verb must has a subject, that "we" is plural, etc. It will learn that "swim" is to "swimming" as "walk" is to "walking", but will not know that "swimming" is a gerund. Note that word2vec can be extended to sentences or whole documents as proposed by Le et al. [3] but this will be out of scope for our research.

1.3 Research Purpose

This research aims to improve the capabilities of machines to learn translations with minimal human intervention. Previous research [6, 8] has already shown potential for word2vec in the context of automatic translation, as discussed in section 1.2.

However, we found no practical implementations using Word2Vec and no further research on different setups for word2vec based machine translation.

2. RESEARCH GOAL

The goal of this research is to give practical advice on a setup for word2vec based machine translation.

Because we aim at a practical purpose of our research, we will also release all code used in our experiments. We hope this makes our research so trivially reproducible that others will further develop and use our results.

2.1 Tasks

To build a test system and perform our experiments, we need to execute the following steps:

- Build (at least) two working setups for automated translation:
  - A single-model setup:
    - Merge datasets, train single model with plain word2vec (easy)
    - Build a translator on top of the model using the same relational structure as in the man/woman king/queen example.
  - A multi-model setup:
    - Train the individual models
    - Train transformation matrix between models, for example using linear regression.
    - Build a translator that converts a word to a vector, "translates" the vector and looks up the corresponding word in the other model.
- Make a verification mechanism
- Verify which setup works better

3. PLANNING

A global overview of the milestones we defined in our research is as follows;

Half September - October: Literary Study)
During this phase, a list of relevant papers (e.g. [4, 6, 8]) is collected, short-listed to a readable size and read. Each paper will get a brief (informal) summarization to capture the essence of the paper insofar as it is relevant to this research. These summaries help process the information and provide quick access during the practical research.

Not all papers relevant to the research will be read during this phase, so next phases will include a fair share of reading.

October: Preparing experiments
In this phase, we gather the required corpora (see Section 5) and implement our translation programs.

November: Running experiments
Running the experiments includes training the models (which might take several days) and running the translation programs against a set of predefined correct translations.

Before 12 December: Incorporating results in paper
During this phase, we will incorporate the results obtained in the experiments into our final paper.

4. ALLOCATION OF RESPONSIBILITIES

Since our datasets are too big to be processed on a laptop, a cluster is needed. This will be provided by SICS (Swedish Institute of Computes Science [1]). Bram will be responsible for downloading and cleaning the different resources of chapter 5 plus uploading these to the cluster. Marc will be responsible for setting up the virtual machines in the cluster in order to run our experiments. This will involve installing the dependencies for word2vec to be able to run it in a distributed mode. When this is done we will run the experiments together.

5. RESOURCES

We have identified four main resources needed for this research: a large English corpus, a large Dutch corpus, a large set of word translations from Dutch to English and computional resources to analyze the data.

The corpora for both languages should be several tens of gigabytes or more. Mikolov et al. [7] specifically state that a large amount of training data is crucial for word2vec to build a correct model. They speak of corpora 30 billion words, with a significant decrease in quality of the results when lowering the corpus size to only 6 billion words.

For this research, we will look at the following corpora:

- *Reddit corpus* [2] (1TB uncompressed JSON, 50 billion words) Corpus containing all posts on Reddit. It is mostly English, but also contains a few other languages.

[1] `http://www.sics.se/`
[2] `https://archive.org/details/2015_reddit_comments_corpus`

- *Wikipedia EN crawl* [3] (60GB uncompressed XML, 3 billion words) Corpus of all text on the English Wikipedia (ignoring revisions).
- *Wikipedia NL Crawl* [4] (5GB uncompressed XML, 250 million words) Corpus of all text on the Dutch Wikipedia (ignoring revisions).
- *SoNaR-500* [5] (60GB uncompressed text, >500 million words) Curated corpus gathered by several Dutch universities, consisting of newspapers, subtitles, etc.
- *Dutch Parallel Corpus* [6] (10 million pairs of words Dutch-English.

These datasets are all open for academic usage, in various degrees of opennes. The Reddit corpus is available through their API or with bittorrent, the Wikipedia data is freely available for download under CC-BY-SA license, and both SoNaR-500 and the Dutch Parallel Corpus are available after signing a special research license agreement.

Before processing the corpora with word2vec, we will convert them to a universal format: plain text, with special characters (everything that is not alphanumeric, space or dash) removed, and one entity (post, article or article) per line.

To make our research reproducible, we will publish all our code online and provide lists with hashes of all our input and intermediate steps (e.g. after cleaning the data).

6. REFERENCES

[1] N. Boycheva. Distributional similarity music recommendations versus spotify: A comparison based on user evaluation. 2015.

[2] C. Buck, K. Heafield, and B. van Ooyen. N-gram counts and language models from the common crawl. In *Proceedings of the Language Resources and Evaluation Conference*, 2014.

[3] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[4] O. Levy, Y. Goldberg, and I. Ramat-Gan. Linguistic regularities in sparse and explicit word representations. *CoNLL-2014*, page 171, 2014.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] L. Wolf, Y. Hanani, K. Bar, and N. Dershowitz. Joint word2vec networks for bilingual semantic representations. *International Journal of Computational Linguistics and Applications*, 5(1):27–44, 2014.

[3] `https://dumps.wikimedia.org/enwiki/20150901/`

[4] `https://dumps.wikimedia.org/nlwiki/20150901/`

[5] `http://tst-centrale.org/producten/corpora/sonar-corpus/6-85`

[6] `http://tst-centrale.org/producten/corpora/dutch-parallel-corpus-niet-commercieel/6-65`