Linguistic Regularities in Continuous Space Word Representations

Tomas Mikolov*, **Wen-tau Yih, Geoffrey Zweig**
Microsoft Research
Redmond, WA 98052

Abstract

Continuous space language models have recently demonstrated outstanding results across a variety of tasks. In this paper, we examine the vector-space word representations that are implicitly learned by the input-layer weights. We find that these representations are surprisingly good at capturing syntactic and semantic regularities in language, and that each relationship is characterized by a relation-specific vector offset. This allows vector-oriented reasoning based on the offsets between words. For example, the male/female relationship is automatically learned, and with the induced vector representations, "King - Man + Woman" results in a vector very close to "Queen." We demonstrate that the word vectors capture syntactic regularities by means of syntactic analogy questions (provided with this paper), and are able to correctly answer almost 40% of the questions. We demonstrate that the word vectors capture semantic regularities by using the vector offset method to answer SemEval-2012 Task 2 questions. Remarkably, this method outperforms the best previous systems.

1 Introduction

A defining feature of neural network language models is their representation of words as high dimensional real valued vectors. In these models (Bengio et al., 2003; Schwenk, 2007; Mikolov et al., 2010), words are converted via a learned lookup-table into real valued vectors which are used as the inputs to a neural network. As pointed out by the original proposers, one of the main advantages of these models is that the distributed representation achieves a level of generalization that is not possible with classical n-gram language models; whereas a n-gram model works in terms of discrete units that have no inherent relationship to one another, a continuous space model works in terms of word vectors where similar words are likely to have similar vectors. Thus, when the model parameters are adjusted in response to a particular word or word-sequence, the improvements will carry over to occurrences of similar words and sequences.

By training a neural network language model, one obtains not just the model itself, but also the learned word representations, which may be used for other, potentially unrelated, tasks. This has been used to good effect, for example in (Collobert and Weston, 2008; Turian et al., 2010) where induced word representations are used with sophisticated classifiers to improve performance in many NLP tasks.

In this work, we find that the learned word representations in fact capture meaningful syntactic and semantic regularities in a very simple way. Specifically, the regularities are observed as constant vector offsets between pairs of words sharing a particular relationship. For example, if we denote the vector for word i as x_i, and focus on the singular/plural relation, we observe that $x_{apple} - x_{apples} \approx x_{car} - x_{cars}$, $x_{family} - x_{families} \approx x_{car} - x_{cars}$, and so on. Perhaps more surprisingly, we find that this is also the case for a variety of semantic relations, as measured by the SemEval 2012 task of measuring relation similarity.

*Currently at Google, Inc.

Proceedings of NAACL-HLT 2013, pages 746–751,
Atlanta, Georgia, 9–14 June 2013.