Word embedding is about learning continuous word representations from corpus. (Mention LSA or LDA for learning continous word representation)

1. CBOW

CBOW use a language model to learn continuous word representations. Its learning target is to maximize the following function.

Or equivalently, maximize

That is, to learn word representations such that these representations enables accurate prediction of the central words from its surround words(context).

The probability p(w|c) is computed using Input word vector of w and output word vector of c

However, this is not practical because of the large sum in the bottom for every word and context pair.

 Ways to improve: Hierarchical softmax. Binary tree of the output layer. Leaves are words in vocabulary. Each inner node is represented by a relative probability. Each inner node n is represented by a vector v\_n’(shared variable), each word w is represented by a vector v\_w. Every time a central word is sampled, .

2. Skip-gram

Skip-gram uses another language model to learn continuous word representations. Learn target is to maximize

Or equivalent, maximize

That is, to learn word representation such that these representation enables accurate prediction of its surrounding words from the central word.

The probability p(w|c) is computed using Input word vector of w and output word vector of c

However, this is not practical because of the large sum in the bottom for every word and context pair.

3. Ways to improve efficiency of training CBOW and Skip-gram