Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection

Richard Socher, Eric H. Huang, Jeffrey Pennington∗ , Andrew Y. Ng, Christopher D. Manning Computer Science Department, Stanford University, Stanford, CA 94305, USA ∗SLAC National Accelerator Laboratory, Stanford University, Stanford, CA 94309, USA richard@socher.org, {ehhuang,jpennin,ang,manning}@stanford.edu

Abstract

Paraphrase detection is the task of examining two sentences and determining whether they have the same meaning. In order to obtain high accuracy on this task, thorough syntactic and semantic analysis of the two statements is needed. We introduce a method for paraphrase detection based on recursive autoencoders (RAE). Our unsupervised RAEs are based on a novel unfolding objective and learn feature vectors for phrases in syntactic trees. These features are used to measure the word- and phrase-wise similarity between two sentences. Since sentences may be of arbitrary length, the resulting matrix of similarity measures is of variable size. We introduce a novel dynamic pooling layer which computes a fixed-sized representation from the variable-sized matrices. The pooled representation is then used as input to a classifier. Our method outperforms other state-of-the-art approaches on the challenging MSRP paraphrase corpus.

Úvod do problému:
Paraphrase detection is the task of examining two sentences and determining whether they have the same meaning.

Upresnenie:
In order to obtain high accuracy on this task, thorough syntactic and semantic analysis of the two statements is needed.

Ako bol problém riešený:
We introduce a method for paraphrase detection based on recursive autoencoders (RAE). Our unsupervised RAEs are based on a novel unfolding objective and learn feature vectors for phrases in syntactic trees. These features are used to measure the word- and phrase-wise similarity between two sentences.

Podproblém, ktorý sa snažia vyriešiť:
Since sentences may be of arbitrary length, the resulting matrix of similarity measures is of variable size.

Riešenie podproblému:
We introduce a novel dynamic pooling layer which computes a fixed-sized representation from the variable-sized matrices. The pooled representation is then used as input to a classifier.

Celkové dosiahnuté výsledky:
Our method outperforms other state-of-the-art approaches on the challenging MSRP paraphrase corpus.