Semantic Representations via Multilingual Sentence Equivalence

Chu-Cheng Lin Kazuya Kawakami

Language Technologies Institute, Carnegie Mellon University

Introduction

- Recently many research have been done in word semantics. In particular, distributed word representation gathers a lot of attention. However we still don't know what is a proper way of representing a sentence.
- In this project, we want to investigate into distributed representations that are constrained by the principle of semantic equivalence.
- We build on the CVM model (Hermann and Blunsom, 2013) and extend the model for multilingual semantic equivalence.
- We evaluate our sentence embeddings on a cross–lingual document classification task.
- Also, we used cross lingual sentence embeddings for reranking translation candidates, we get marginally better results on a small test set.

Multilinguality

Multilingual Data will add information such as

- Grammatical Difference
- Synonym / Antonym
- Wordsense Disambiguation (e.g. bank is translated to "banque" or "rive")

CVM

- The CVM model just adds individual word embeddings to get the sentence embeddings.
- We want to make v_e accord with v_f.
- Specifically, we use this loss function:

$$cost(v_e, v_f, v', v'') = 2\log(dist(v_e, v_f)) - \log(dist(v_e, v')) - \log(dist(v_f, v'')), \quad (1)$$

where v' and v'' are embeddings of sentences that are known to be different from v_e and v_f.

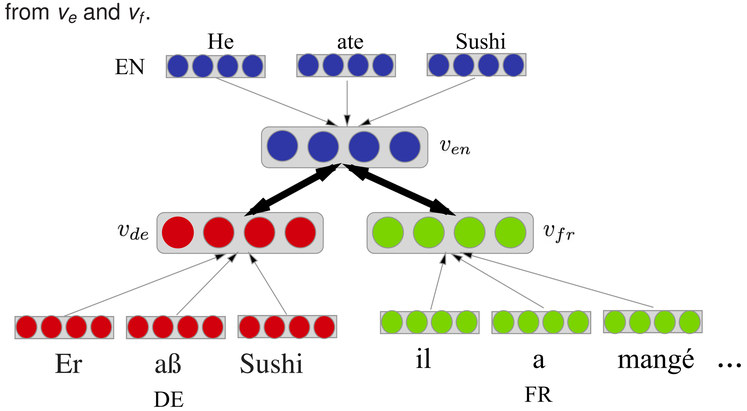

Figure: CVM

Cross lingual document classification

We evaluate our CVM sentence embeddings on a German-English cross-lingual document classification task. In this task we train a perceptron classifier using sentence embeddings of German texts, and classify test English documents.

We want to know

- how much does multilinguality help: we test with these language settings: $(de, en), (de, fr, en), (de, fi, en), (de, cs, en), (de, zh, en)$
- if initialization with existing monolingual embeddings help: we compare between two initialization strategies
  - glove embeddings
  - random vectors

We use Europarl except for Chinese throughout the experiments. For each language we use 100000 translation pairs. Optimization is done with AdaDelta.

Reranking translation candidates

- First we train a German–English MT system with MOSES and obtain features $f_1 \dots f_N$ and associated weights $w_1 \dots w_N$.
- We add a new feature $d_{s,c}$, the Euclidean distance of translation candidate e_c to the source sentence s.
- Fixing $w_1 \dots w_N$, we search for optimal weight w_d that maximizes BLEU on the test set.
- Results are in Tab 2.

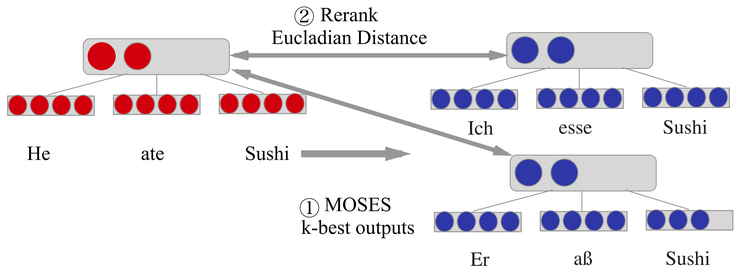

Figure: Reranking Task

References

K. Hermann and P. Blunsom
Multilingual distributed representations without word alignment.
arXiv preprint arXiv:1312.6173, 2013.

Results

Cross Lingual Document Classicication Task (RCV1/RCV)
Train Classifier with German sentence vector, and Test with English sentence vector.

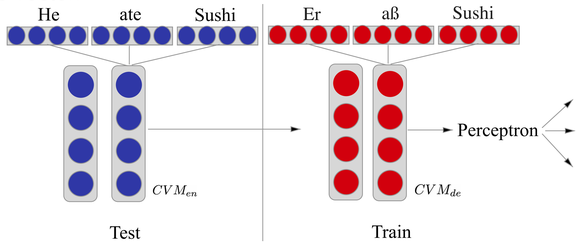

Figure: RCV1/RCV2 Task Description

init strategy (de)	$+en$	$+fr+en$	$+fi+en$	$+cs+en$	$+zh+en$	$+fr+cs+en$
random	0.436	—	0.467	—	0.467	—
glove	0.661	0.664	0.673	0.676	0.614	0.685

Table: Results from CVMs on the CLDC task.

Metric	BLEU	METEOR	TER
original score	32.7	51.5	51.8
reranked score	32.8	51.4	51.6

Table: Results from reranking

Discussion

- Language pair: adding Finnish, Czech and French helps accuracy. But adding *initialized* Chinese actually hurts accuracy. However the best results are obtained from a run with both French and Czech.
- Loss function: standard hinge loss doesn't do well.
- Initializing with GloVe embeddings helps a lot. However adding extra uninitialized languages help.
- The quality of word embeddings obtained from CVM models are not suitable for word similarity tasks (WS-353 dataset).

Future Work

- We tried LSTM for sequence to sequence translation models but they suffered high perplexity. Probably performance can be improved with more data.
- In the future we hope to generate sentences from these learned sentence embeddings.