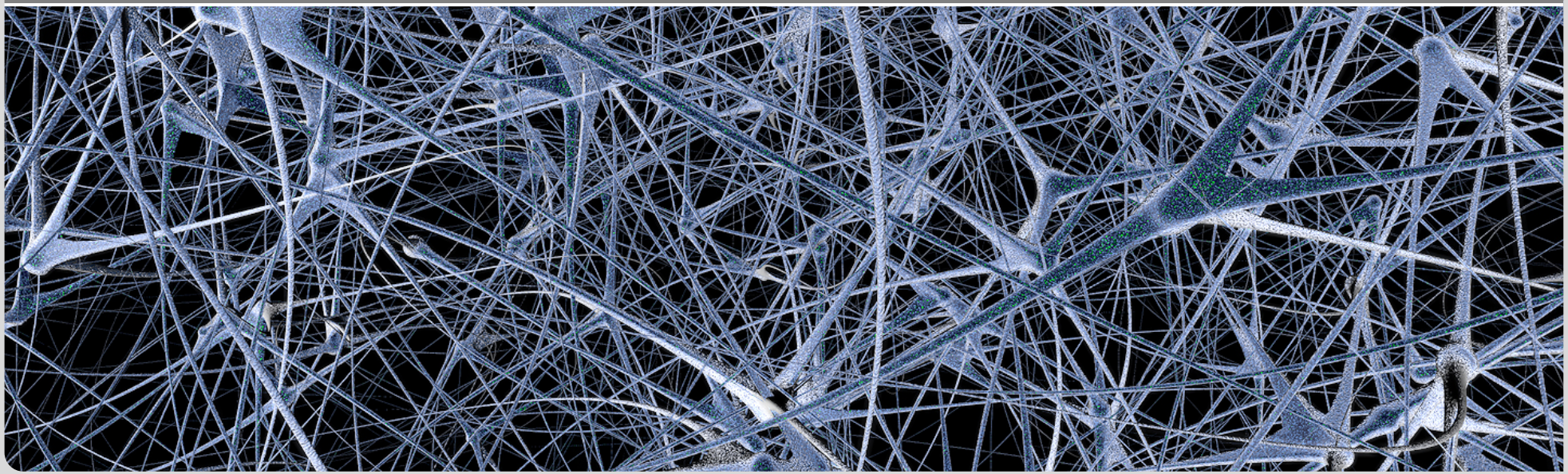
Karlsruhe Institute of Technology

Lexical Translation Model using a Deep Neural Network Architecture

Thanh-Le Ha, Jan Niehues and Alex Waibel

Institute for Anthropomatics and Robotics, Interactive Systems Labs



Phrase-based SMT – Context problems

German: Der Mörder bringt sein kleines Opfer heimlich im Parkhaus um
Correct: In the parking garage, the murderer kills his victim secretly
Wrong: The murderer brings the victim secretly in the parking garage around

- Translation model: Can only perform the lexical translation within phrases
  - Phrases are locally determined as they are word sequences
  - Cannot model long-distance translation without reordering
- Language model: A context up to *n* words can be considered only

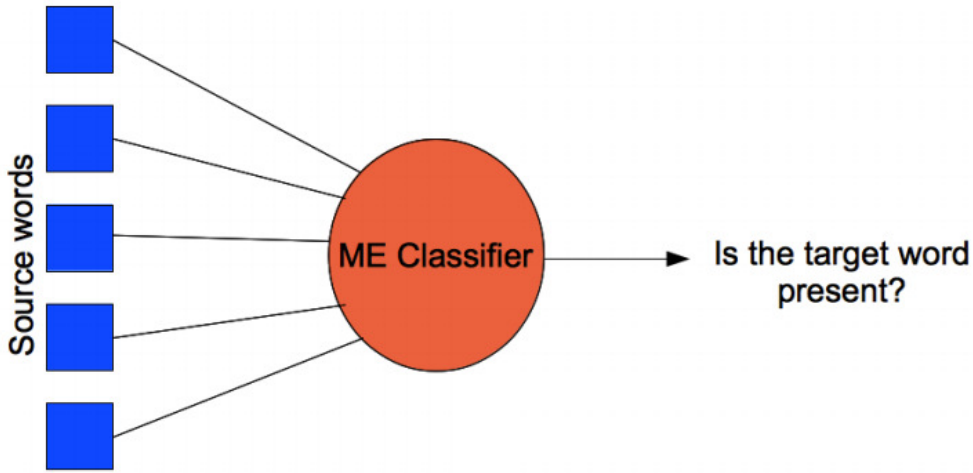
Motivation

- Exploit global contexts beyond the phrasal boundaries
- Maintain local contexts against bag-of-word models
- Learn shared word representation space:
  - To better model dependencies between source words
  - To perform abstraction over discrete word representation

Related work

- Utilize global contexts:
  - Treat the lexical selection as a WSD problem (Carpuat et al., 2007)
  - Predict appearances of target words based on the source words (Bangalore et al., 2007, Mauser et al., 2009, Niehues and Waibel, 2013)
- Perform translation over shared word representation using deep NNs:
  - Model target word's probability given history of bilingual n-grams (Le et al., 2012, Devlin et al., 2014)
  - Calculate probability of target phrases given source phrases (Schwenk et al., 2012)

Discriminative Word Lexicon (DWL)

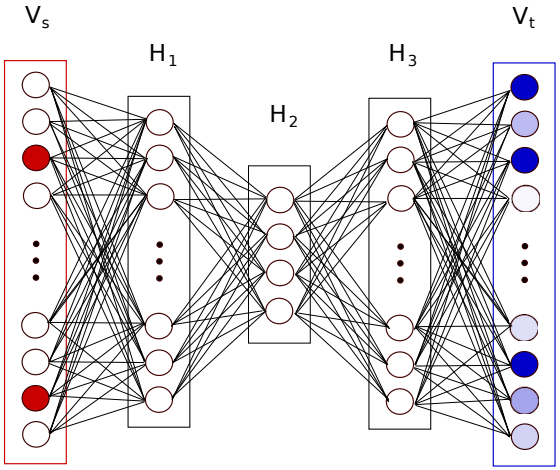


- Original idea from (Mauser et al., 2009)
  - Build a classifier for every target word, given source words' presences
  - Combine the probabilities of target words to form the sentence score.
- Extensions from (Niehues and Waibel, 2013)
  - Extend DWL to employ local contexts (source n-grams, target surrounding words)
- Both use Maximum Entropy (ME).

DWL using Deep Neural Network - Idea

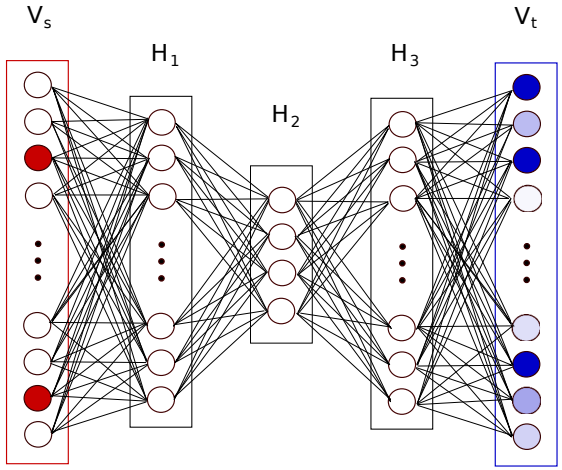
- Introduce the non-linearity into the DWL
  - Make predictions over shared word representation of the source side
- Better model the dependencies among features
  - Features in DWL could be extended to n-grams, word classes, POS tags, ..

DWL using Deep Neural Network - Architecture



- NNDWL: A Feed-forward neural network with 3 hidden layers H_1, H_2 and H_3
- Input: the bag-of-word representation (BoW) of source sentence s, size V_s
- Output: Probabilistic decision for the presence of all target words, size V_t
- Activation function: sigmoid (multivariate binary classifiers)

DWL using Deep Neural Network - Training



- Calculate the output of the network with current weights (forward pass).
- Back-propagate Cross Entropy error between $\hat{t}$ and $\vec{p}$ to update the weights:

$$E = -\frac{1}{V_t} \sum_{i=1}^{V_t} (\hat{t}_i \ln p_i + (1 - \hat{t}_i) \ln(1 - p_i))$$

$\hat{t}$: BoW vector of the training target sentence
$\vec{p} = [p_i]$: Output of the network

Practical Issues

- Difficulty: Training is considerably expensive, due to:
  - Non-linearity over hidden layers
  - Calculations of the output over the whole target vocabulary
- Solutions:
  - Use rectified linear units on hidden layers
  - Model the DWL for the k most frequent words only
  - Others are treated as unknown words
- Utilize sparse representation for the input
- Try drop-out as a regularization

Evaluation - System Description

- TED translation task: IWSLT English→French (EN-FR) 2013
  - Baseline: The best system for IWLST EN-FR 2013 without MEDWL:
    - Phrase-based MT, train on large training data
    - PT and LM Adaptations
    - Bilingual, POS-based and cluster language models
    - Lexicalized reordering
  - DWL only trained on the TED corpus

Evaluation - Different vocabulary sizes

- The size of the source and target sides varies in $\{500, 1000, 2000, 5000\}$

System (En-Fr)	BLEU	ΔBLEU
Baseline	31.94	–
MaxEnt DWL	32.17	+0.23
NNDWL 500	32.06	+0.12
NNDWL 1000	32.37	+0.43
NNDWL 2000	**32.38**	**+0.44**
NNDWL 5000	32.07	+0.13
Full NNDWL	32.06	+0.12

Table : *Results of the English→French NNDWL.*

Evaluation - Impact of source contexts

System (En-Fr)	BLEU	ΔBLEU
Baseline	31.94	–
NNDWL 2000	**32.38**	**+0.44**
NNDWL 2000 SC-200-100	32.35	+0.41
NNDWL 2000 SC-500-200	**32.44**	**+0.50**
NNDWL 2000 SC-1000-500	32.36	+0.42

Table : *Results of the 2000-NNDWL with source contexts.*

System (En-Fr)	BLEU	ΔBLEU
Baseline	31.94	–
NNDWL 1000	**32.37**	**+0.43**
NNDWL 1000 SC-200-100	32.01	+0.07
NNDWL 1000 SC-500-200	32.23	+0.29
NNDWL 1000 SC-1000-500	**32.39**	**+0.45**

Table : *Results of the 1000-NNDWL with source contexts.*

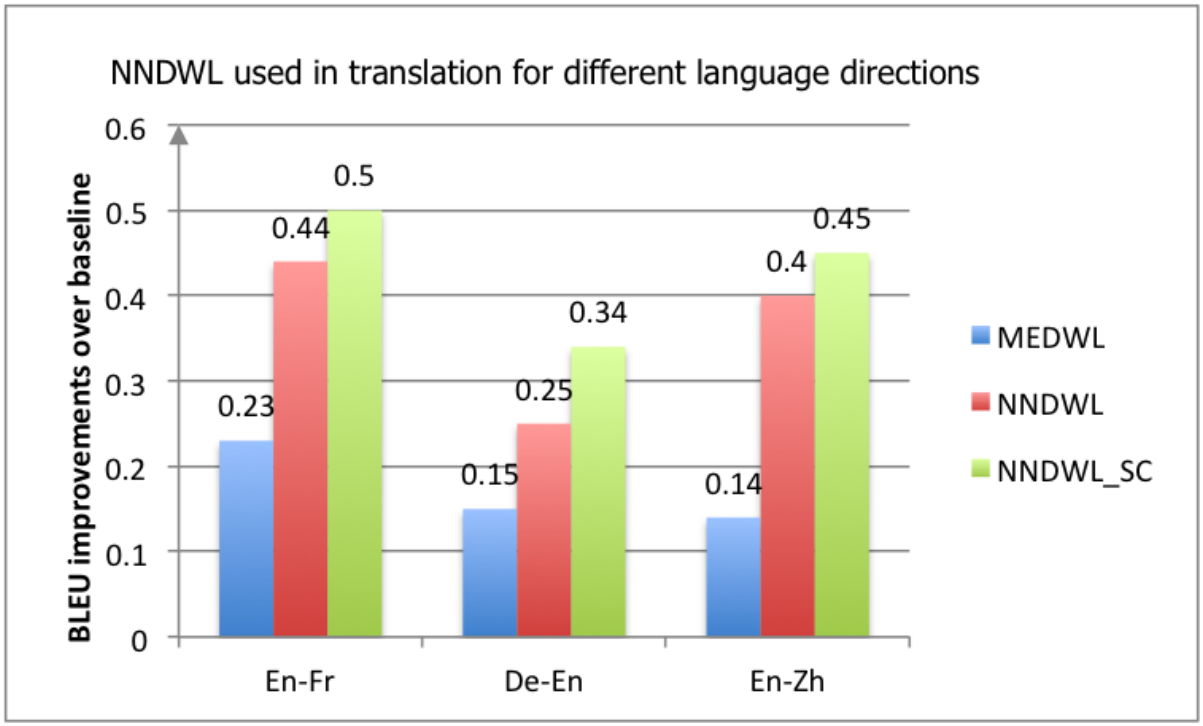
Evaluation - Different architectures

- SimNNDWL: A neural network with one hidden layer.

System (En-Fr)	BLEU	ΔBLEU
Baseline	31.94	–
NNDWL 1000	**32.37**	**+0.43**
SimNNDWL 1000	32.12	+0.18
NNDWL 2000	**32.38**	**+0.44**
SimNNDWL 2000	32.29	+0.35

Table : *Results of NNDWL and SimNNDWL architectures.*

Evaluation - Different language pairs



Conclusion & Future work

Conclusion

- DWL using deep neural networks improves over not using it or using the original one, on different language pairs.
- Deep architecture works better than the simpler ones.
- The configuration of the network need to be chosen carefully.

Future work

- Evaluate the accuracy of lexical selection methods
- Integrate other non-word features
- Conduct more experiments with other configurations

Thank you

DWL example

ME-model for *teacher*

Source: der Lehrer redet .

Features: .=1; der=1; Lehrer=1; redet=1

Target: the teacher talks .

Label: 1

DWL example

ME-model for *teacher*

Source: die Lehrer reden .

Features: .=1; die=1; Lehrer=1; reden=1

Target: the teachers talk .

Label: 0

DWL example

ME-model for *teacher*

Source: das Auto fährt .

Features: .=1; Auto=1; das=1; fährt=1

Target: the car drives .

Label: 0

Source-context DWL

- Original DWL is a bag-of-word model:

Source: Die Lehrer fragten den Schüler Die Schüler fragten den Lehrer
Target: The teachers ask the student The students ask the teacher

↓

Features: die=1;den=1;fragten=1;Lehrer=1;Schüler=1

- Employ source contexts in DWL:

Source: die Lehrer wussten nicht, ...

↓

Features: ,=1; die=1; Lehrer=1; nicht=1; wussten=1;
die_Lehrer = 1; Lehrer_wussten=1;...

Source-context DWL

- Original DWL is a bag-of-word model:

Source: Die Lehrer fragten den Schüler Die Schüler fragten den Lehrer
Target: The teachers ask the student The students ask the teacher

↓

Features: die=1;den=1;fragten=1;Lehrer=1;Schüler=1

- Employ source contexts in DWL:

Source: die Lehrer wussten nicht, ...

↓

Features: ,=1; die=1; Lehrer=1; nicht=1; wussten=1;
die_Lehrer = 1; Lehrer_wussten=1;...

Sentence Probability

- With the independence assumption of target words:

$$p(t|s) \approx \prod_{j=1}^{J} p(t_j|s)$$

J: Length of the target sentence t

- Then the score of every phrase pair can be calculated before translation