# Vector Space Models for Phrase-based Machine Translation

**Tamer Alkhouli, Andreas Guta, and Hermann Ney**

<surname>@cs.rwth-aachen.de

**Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation**
**Doha, Qatar**
**October 25, 2014**

**Human Language Technology and Pattern Recognition**
**Chair of Computer Science 6**
**Computer Science Department**
**RWTH Aachen University, Germany**

# Outline

- **Introduction and Motivation**
- **From Words to Phrases**
- **Semantic Phrase Features**
- **Paraphrasing and Out-of-vocabulary Reduction**
- **Experiments**
- **Conclusion**

# Introduction and Motivation

- **Goal: improve phrase-based translation (PBT) using vector space models**

- **Categorical word representations: no information about word identities**
- **Embedding words in a vector space allow such encoding**
  - **geometric arrangements in the vector space**
  - **enables information retrieval approaches using a similarity measure**

- **Distributional hypothesis (Harris 1954): words occurring in similar contexts have similar meanings**

- **Word representations based on:**
  - **co-occurrence counts (Lund and Burgess, 1996; Landauer and Dumais, 1997) $\rightarrow$ dimensionality reduction (e.g. SVD)**
  - **neural networks (NN) $\rightarrow$ input/output weights**

# From Words to Phrases

- How to learn phrase vectors?
- Phrase representations
  - decompositional approach: resort to word constituents (Gao et al., 2013; Chen et al., 2010)
  - atomic treatment of phrases (Mikolov et al., 2013b; Hu et al., 2014)
    - advantage: reuse word-level methods
    - challenge: data sparsity
- This work: NN-based atomic phrase representations

# Phrase Corpus

- **Phrase corpus** used to learn phrase vectors
- **Corpus built using a multi-pass greedy algorithm**
  - initialization: phrases have length 1
  - join phrases forwards, backwards or do not join
  - Use **bilingual phrase table scores** to make the decision:

$$score(\tilde{f}) = \max_{\tilde{e}} \left\{ \sum_{l=1}^{L} w_l g_l(\tilde{f}, \tilde{e}) \right\}$$

    - $(\tilde{f}, \tilde{e})$: bilingual phrase pair
    - $g_l(\tilde{f}, \tilde{e})$: $l$-th feature of the bilingual phrase pair
    - $w_l$: $l$-th feature weight
- **2 phrasal and 2 lexical features with manually tuned weights**

# Semantic Phrase Feature

- **Add a vector-based feature to the log-linear framework of PBT:**

$$h(\tilde{f},\tilde{e}) = sim(Wx_{\tilde{f}}, z_{\tilde{e}})$$

  - $x_{\tilde{f}}$: $S$-**dimensional source phrase vector**
  - $z_{\tilde{e}}$: $T$-**dimensional target phrase vector**
  - $W$: $T \times S$ **linear projection matrix (Mikolov et al. 2013a)**
  - $sim$: **similarity function (e.g. cosine similarity)**

- **Learn $W$ using stochastic gradient descent**

$$\min_W \sum_{n=1}^{N} ||Wx_n - z_n||^2$$

**where** $(x_n, z_n) \mathrel{\widehat{=}} (x_{\tilde{f}}, z_{\tilde{e}})$ **such that:**

$$\tilde{e} = \operatorname*{argmax}_{\tilde{e}'} \left\{ \sum_{l=1}^{L} w_l g_l(\tilde{f}, \tilde{e}') \right\}$$

# Out-of-vocabulary Reduction

- **Introduce new phrase pairs to the phrase table**
- **Paraphrase $\tilde{f}$ with $|\tilde{f}| = 1$**
  - **reduce out-of-vocabulary (OOV) words**
  - **use word vectors**
- **$k$-nearest neighbor search using a similarity measure**

- **Additional phrase table feature**
  - **similarity measured between a phrase and its paraphrase**
  - **original features copied from original phrase pair**
- **Avoid interfering with existing phrase entries**

  $\rightarrow$ **limit paraphrasing to source words unseen in parallel data**

# Experiments

- **IWSLT 2013 Arabic→English task**
- **Domain: TED lectures**

| | TED | | UN | |
|---|---|---|---|---|
| | **Arabic** | **English** | **Arabic** | **English** |
| Sentences | 147K | | 8M | |
| Running Words | 3M | 3M | 228M | 226M |

**IWSLT 2013 Arabic and English corpora statistics**

# Experiments

- **Phrase vectors trained using *word2vec*[1]**
  - **simple neural network model without hidden layers**
  - **use frequent phrases only**

- **Vector dimension: Arabic: 800, English: 200**
- **5 passes for phrase corpus construction**

[1] http://code.google.com/p/word2vec/

# Experiments

| | TED+UN | |
| --- | --- | --- |
| | **Arabic** | **English** |
| # tokens | | |
| words | 231M | 229M |
| phrases | 126M | 115M |
| vocabulary | | |
| words | 0.5M | 0.4M |
| phrases | 5.8M | 5.3M |
| # vectors (*word2vec* vocabulary) | | |
| words | 134K | 123K |
| phrases | 934K | 913K |

**Corpus and vector statistics for IWSLT 2013 Arabic→English**

# Experiments

- Standard PBT Baseline features:
  - 2 phrasal features
  - 2 lexical features
  - 3 binary count features
  - 6 Hierarchical reordering features
  - 4-gram mixture LM
  - jump distortion
  - phrase and word penalties

- In-domain baseline data: TED
- Full baseline data: TED+UN, domain-adapted phrase table

# Experiments

- **Word vectors used for paraphrasing**
- **Reduction of OOV rate:** $5.4\% \rightarrow 3.9\%$

| | Arabic | |
|---|---|---|
| | **dev** | **eval13** |
| | # OOV | |
| TED | 185 | 254 |
| TED+paraphrasing | 150 | 183 |
| Vocabulary | 3,714 | 4,734 |

**OOV reduction for IWSLT 2013 Arabic→English**

# Experiments

- **Improvements over the TED baseline**
  - **semantic feature:** $0.4\%$ **BLEU and** $0.7\%$ **TER**
  - **paraphrasing:** $0.6\%$ **BLEU and** $0.7\%$ **TER**

| | **dev2010** | | **eval2013** | |
|---|---|---|---|---|
| **system** | BLEU [%] | TER [%] | BLEU [%] | TER [%] |
| **TED** | 29.1 | 50.5 | 28.9 | 52.5 |
| + semantic feature | 29.1 | †**50.1** | †29.3 | †**51.8** |
| + paraphrasing | 29.2 | †50.2 | †**29.5** | †**51.8** |
| + both | 29.2 | 50.2 | †29.4 | †**51.8** |
| **TED+UN** | 29.7 | 49.3 | 30.5 | 50.5 |
| + semantic feature | 29.8 | 49.2 | 30.2 | 50.7 |

**Semantic feature and paraphrasing results for IWSLT 2013 Arabic → English.**

- **†: statistical significance with** $p < 0.01$

# Conclusion

- **Improved end-to-end translation using vector space models**
  - **semantic phrase features using phrase vectors**
  - **paraphrasing using word vectors**

- **Exploit monolingual data for OOV reduction**

- **Proposed methods helpful for resource-limited tasks**
- **BLEU and TER may underestimate semantic models**

# Thank you for your attention

**Tamer Alkhouli**
**Andreas Guta**

<surname>@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/