Karlsruhe Institute of Technology

Domain Adaptation in SMT using Factored Translation Models

Jan Niehues and Alex Waibel

Institute for Anthropomatics – Prof. Waibel

Overview

- Motivation
- Related Work
- Factored Domain Model
  - Domain Factors Translation Model
  - Domain Factors Sequence Model
- Evaluation
  - News Task
  - Lecture Task
- Conclusion

Motivation

- Large amounts of training data are needed for SMT systems
- Best to have data from similar topics and genre
  - Possible only for few scenarios
    - European Parliament
  - Not possible for many real-world scenarios
  - Example:
    - Lecture Translation
    - Even News for some languages
- Common technique:
  - Use all available data to build a baseline system
  - Adapt system using in-domain data

Motivation

- How to adapt the system?

- State-of-the-art SMT Systems:
  - Assumption: All training sentences are equally important
  - No longer holds if we have in-domain and out-of-domain data
  - Leads especially to many errors if in-domain data is small

- In-domain data should be more important
  - Introduce sentence weights into SMT model

Motivation

- Model domain of the training data explicitly
  - Integrate corpus identifier into the translation model
- Prefer phrase pairs learned from in-domain data
  - Weights can be tuned automatically
- Integration using Factored Translation Models (Koehn and Hoang (2007))
  - Easy to integrate into state-of-the-art SMT systems

Related Work

- Only monolingual in-domain data
  - Language Model Adaption
    - Inspired by work, that was done for ASR
  - Creating synthetic parallel text
    - Translate monolingual text using a baseline system
    - Use translated text as additional training data
    - Ueffing et al. 2007, Schwenk and Senellart 2009

Related Work

- Only Monolingual in-domain data

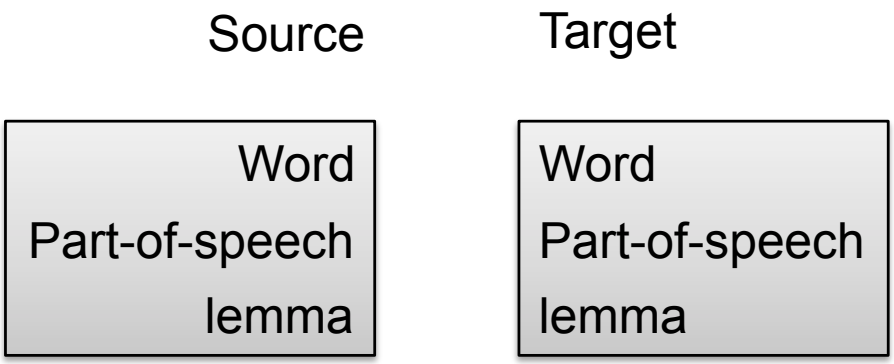
- Parallel in-domain data
  - Combine translation models using alternate decoding paths
    - (Koehn and Schroeder (2007))
  - Adapt translation models using mixture models
    - (Foster and Kuhn (2007))
    - Linear and log-linear combination
    - Different methods to set weights for domain
  - Discriminative weights for sentences of parallel corpus
    - (Matsoukas et at. 2009)

Related Work

- Only Monolingual in-domain data
- Parallel in-domain data
- Data selection
  - Select similar sentences using (cross-lingual) information retrieval techniques
    - Hildebrand et al. 2005
    - Snover et al. 2008

Factored Translation Model

- Framework to integrate corpus id
- Represent words by vector of factors instead of token
  - Integrate additional annotation into SMT



- mainly used to incorporate additional linguistic knowledge

Factored Domain Model

- Model Domain of data explicitly
  - Directly model influence of in-domain and out-of-domain data
  - Optimize weights on development data
- Representation of domain
  - Introduce corpus identifier for different training sources

Bogen # arc # NEWS

Bogen # sheet # EPPS
Bogen # arc # EPPS

- Assumption: Phrase pairs extracted from in-domain data are more important

Factored Domain Model

- Integration into SMT system:
  - Use corpus id as an additional target factor

Ein	blauer	Bogen	(demokratischer)	Staaten im Osten
A IN	blue OUT	arc IN	(democratic) IN IN IN	states in the east IN IN IN IN

- Translation differ by generated target words and domains of these words

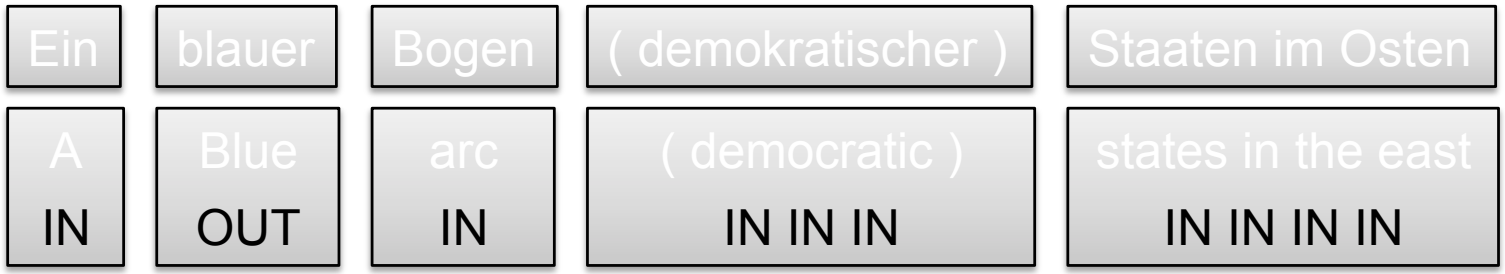
Factored Domain Model

- Domain representation during translation
  - Sequence of corpus ids
  - Example:

Ein	blauer	Bogen	(demokratischer)	Staaten im Osten
A	blue	arc	(democratic)	states in the east
IN	OUT	IN	IN IN IN	IN IN IN IN

Factored Domain Model

- Domain representation during translation
  - Sequence of corpus ids
  - Example:



- Assumption: Prefer translation with more corpus ids from text similar to test domain

Factored Domain Model

- Phrase pairs that occur in both corpus
  - Example:
    - Bogen # arc # IN
    - Bogen # arc # OUT
- Lead to different phrase pair
  - Existing phrase pair scores are the same
  - New model scores are different
- Select best one according to current weights

Factored Domain Model

- Describe probability of the domain by two additional models

- Domain Factors Translation Model:
  - Probability of generating a sequence of corpus id tags given the sentence
  - Example:
    - im Osten # in Eastern Europe # IN -> quite low probability
    - im Osten # in Eastern Europe # OUT -> higher probability

- Domain Factors Sequence Model:
  - How probable is a sequence of corpus id tags
  - Example:
    - IN OUT IN IN IN IN IN IN IN IN -> high probability
    - IN OUT OUT IN IN IN OUT OUT OUT OUT -> lower probability

Domain Factors Translation Model

- Probability of generating a sequence of corpus id tags given the sentence
  - Similar to phrase translation model in state-of-the-art SMT approach
  - Modeled using 2 two scores

$$P(t \mid s) = \frac{cooc(s,t)}{cooc(s,*)}$$

$$P(s \mid t) = \frac{cooc(s,t)}{cooc(*,t)}$$

  - Estimated using cooccurence counts *cooc(s,t)*

Domain Factors Translation Model

- Cooccurrence count depending on three parameters
  - Use *cooc(s,t,d)* instead of *cooc(s,t)*
- Leads to 3 different probabilities
- Domain Frequency:
  - Probability of the domain tags given the phrase pair
  - Can be approximated by:

$$P(d \mid s,t) = \frac{cooc(s,t,d)}{cooc(s,t,*)}$$

Domain Factors Translation Model

- Target Frequency:
  - Probability of the target phrase given source and domain sequence

$$P(t \mid s,d) = \frac{cooc(s,t,d)}{cooc(s,*,d)}$$

  - Equal to:
    - Target Translation Probability restricted to phrase pairs extracted only from the one domain

- Source Frequency:
  - Probability of the source phrase given target and domain sequence

$$P(s \mid t,d) = \frac{cooc(s,t,d)}{cooc(*,t,d)}$$

Domain Factors Sequence Model

- Describe probability of a sequence of corpus id tags
- Similar to language model
- Problem: cannot be trained on training data
  - Training sentences are always from one domain
- Use discriminative uni-gram model

Domain Factors Sequence Model

- Word Count model
  - Number of target words translated by in-domain phrase pairs
- Phrase Count model
  - Number of phrase pairs extracted from the in-domain corpus
- Several different corpora:
  - One feature for every corpus id
  - Example:
    - ID1 ID2 ID1 ID3 ID1 ID3 -> (3 1 2)

Evaluation

- Two task for translating from German to English
  - Translation of News-commentary texts
    - Test set: News test set of WMT 2007
    - Out-of-domain Data: European Parliament (ca. 33 M Words)
    - In-domain Data: News commentary (ca. 1 M Words)
  - Translation of university Lectures
    - Test set: Different lectures
    - Data: European Parliament, BTEC and News Commentary
    - In-domain Data: Lecture 200K Words

Evaluation

- Preprocessing:
  - Normalization of different German writing systems
  - Compound Splitting
- Discriminative Word Alignment
- Phrase extraction according to moses scripts
  - Additional smoothing of relative frequencies
- POS-based reordering for short and long-range reorderings
  - Rules learned from word-aligned corpus
  - Different reorderings encoded in lattice

Domain Factors Sequence Models

- News translation task
- Domain Factors Translation Model:
  - Domain Relative Frequencies

	System	Dev	Test
1	Baseline	25.90	29.03
2	(1) + LM Adaptation	26.68	29.24
3	(2) + Domain Rel. Frequency	26.80	29.21
4	(3) + Word Count Model	27.03	29.63
5	(3) + Phrase Count Model	27.09	29.54

Domain Factors Translation Models

- News translation task
- Domain Factors Sequence Model:
  - Word Count Model

	System	Dev	Test
1	Baseline	25.90	29.03
2	(1) + LM Adaptation	26.68	29.24
3	(2) + Word Count Model	26.13	29.17
4	(3) + Domain Frequency	27.03	29.63
5	(3) + Target Frequency	27.00	29.51
6	(3) + Source Frequency	26.95	29.84
7	(3) + All	27.07	29.69

Lecture Task

- Domain Factors Sequence Model:
  - Word Count Model

	System	Dev	Test
1	LM Adaptation	36.93	29.84
2	(1) + Source Frequency	37.90	31.12
3	(1) + Target Frequency	37.63	30.73
4	(1) + Domain Frequency	37.28	30.16
5	(1) + All	37.74	31.53
6	(2) + All Sep. corpus ids	38.01	31.51

Example Translations

- Input:
  - Ein blauer Bogen (demokratischer) Staaten im Osten, ...
- Reference:
  - An arc of blue (Democratic) states in the East, ...
- Baseline:
  - A blue sheet (democratic) countries in Eastern Europe, …
- Adapted:
  - A blue arc (democratic) states in the east, …

Conclusion

- New approach to adapt phrase-based SMT systems using Factored Translation Models
  - Easy to integrate
- Model domain of training corpus explicitly
  - Introduce corpus id
  - Add two types of features to log-linear model
  - Weights can be optimized using MERT
- Translation performance could be improved by up to 1 BLEU point