International Workshop on Spoken Language Translation
Kyoto, Japan
September 30 - October 1, 2004

Statistical Machine Translation of Spontaneous Speech with Scarce Resources

Evgeny Matusov, Maja Popović, Richard Zens, and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University
D-52056 Aachen

Content

1. overview: data sparseness problem
2. overview: statistical machine translation
3. acquiring additional training data
4. morphological information for word alignments
   - lexicon smoothing
   - hierarchical lexicon counts
5. part-of-speech information for reordering
6. experimental results
7. summary and outlook

Overview: Translation with Scarce Resources

- language pair specific data sparseness
- lack of bilingual sentence-aligned data in a specific domain (e.g. spontaneous utterances)
- limited coverage of the vocabulary (e.g. highly inflected languages)
- insufficient data to learn non-monotonous translations

Related work

- S. Nießen and H. Ney. 2001. Morpho-syntactic analysis for Reordering in Statistical Machine Translation. In Proc. MT Summit VIII, pages 247–252, Santiago de Compostela, Galicia, Spain, September.
- S. Nießen and H. Ney. Toward hierarchical models for statistical machine translation of inflected languages. In *Data-Driven Machine Translation Workshop*, pages 47–54, Toulouse, France, July.
- F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.
- D. Sündermann and H. Ney. 2003. Synther – a new m-gram POS tagger. In *Proc. NLP-KE-2003, International Conference on Natural Language Processing and Knowledge Engineering*, pages 628–633, Beijing, China, October.
- R. Zens and E. Matusov and H. Ney. 2004. Improved Word Alignment Using a Symmetric Lexicon Model. In *Proc. COLING04*, pages 36–42, Geneva, Switzerland, August.
- Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, and K. Yamada. 2002. Translating with Scarce Bilingual Resources. *Machine Translation* 17, pp. 1–17.

Overview: Statistical Machine Translation

- **source string $f_1^J = f_1 ... f_j ... f_J$ to be translated into a target string $e_1^I = e_1 ... e_i ... e_I$.**
- **classical source-channel approach:**

$$\begin{aligned} \hat{e}_1^I &= \operatorname*{argmax}_{e_1^I} \left\{ Pr(e_1^I | f_1^J) \right\} \\ &= \operatorname*{argmax}_{e_1^I} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} \end{aligned}$$

- $Pr(e_1^I)$: **language model**
- $Pr(f_1^J | e_1^I)$: **translation model**
- **word alignment is introduced as a hidden variable:**

$$Pr(f_1^J | e_1^I) = \sum_A Pr(f_1^J, A | e_1^I)$$

Statistical word alignments

- **alignment A is a mapping from source sentence positions to target sentence positions $a_1 ... a_J$, $a_j \in \{0, \ldots, I\}$.**
- **alignment may contain connections $a_j = 0$ with the 'empty' word e_0**
- **commonly used translation models: IBM-1 to IBM-5, HMM.**
- **all of the models include single-word based lexicon parameters $p(f|e)$**
- **model parameters are trained iteratively with the EM algorithm**
- **usually: restricted alignments (many-to-one mappings only), alignment combination heuristics**
- **recent suggestions: symmetrized lexicon models, symmetric alignments (Zens, Matusov, Ney: *CoLing* 2004)**

Translation

- primary model: alignment templates
  - pairs of source and target phrases and the alignment within the phrases
  - extracted from word alignments
  - automatically trained word classes are used instead of words for better generalization
- search: direct modeling of the posterior probability $Pr(e_1^I | f_1^J)$ using a loglinear model
- easy integration of additional models/feature functions
  - word translation model
  - a word trigram and a class-based five-gram language model
  - word penalty, alignment template penalty, ...
- minimum error training of model scaling factors

Acquiring Additional Training Data

- include additional bilingual training data from other sources
- select domain-relevant data only
- relevance measure: n-gram coverage
- compute the set C of n-grams occuring in the source part of the initial (small) training corpus
- count the occurrence of the n-grams from C in the additional sentences
- coverage score: geometric mean of n-gram precisions ($n = 1, 2, 3, ..., 4$)
- add only sentences with high coverage score

Morphological Information for Word Alignments

- **common statistical lexicon models are based on full form words only**
- **lexicon coverage is low, especially when training with scarce data**
- **a big problem for highly inflected languages like German**
- **smooth the lexicon model with a backing-off lexicon based on word base forms**
- **perform smoothing after each iteration of the EM algorithm**
- **smoothing technique: absolute discounting with interpolation:**

$$p(f|e) = \frac{\max\{N(f,e) - d, 0\}}{N(e)} + \alpha(e) \cdot \beta(f|\overline{e})$$

- $\overline{e}$ **is the base form (generalization) of** e**.**
- **backing-off distribution:** $\beta(f|\overline{e}) = \frac{N(f,\overline{e})}{\sum\limits_{f'} N(f',\overline{e})}$

Hierarchical Lexicon Counts

- **for each German word, determine the base form and sequence of morpho-syntactic tags**
  - **e.g. gehe#gehen-V-IND-PRES#gehen**
- **collect three types of counts in the E-step of the EM algorithm:**
  - **regular full form counts** $N(f, e)$
  - **base form+tag counts** $N(\tilde{f}, e)$
  - **base form counts** $N(\overline{f}, e)$
- ***in each iteration*, combine these counts to hierarchical counts:**

$$N_{hier}(f, e) \;=\; N(f, e) + N(\tilde{f}, e) + N(\overline{f}, e)$$

- **M-step: obtain new estimation of the lexicon probability:**

$$p(f|e) \;=\; \frac{N_{hier}(f, e)}{\sum\limits_{f'} N_{hier}(f', e)}$$

Monotonization of Translation Process

- some language pairs have significantly different word order
- with limited training data, word alignments and phrase structures are estimated poorly
- differences in word order can be reduced by re-ordering of the source sentences (in training and in testing)
- re-ordering rules: using part-of-speech information and knowledge about target sentence structure
- POS tags obtained by using a statistical POS tagger
- POS information is less context-dependent than a syntactic tree structure and thus can be relied upon even when tagging spontaneous utterances
- monotonization of alignments will result in more robust phrase extraction (e.g. non-contiguous phrases can be extracted)

Reordering Rules - 1

- **verb prefixes:**

```
  Ich fahre um 9 Uhr vom Bahnhof ab
> Ich fahre ab um 9 Uhr vom Bahnhof
```

- **compound verbs:**

```
  Ich kann Ihnen noch heute meine Nummer geben
> Ich kann geben Ihnen noch heute meine Nummer
```

- **verb position in subordinate clauses:**

```
  ...  weil ich erst dann Ihnen meine Nummer geben kann
> ...  weil ich kann geben erst dann Ihnen meine Nummer
```

Reordering Rules - 2

- translation improvements:

```
  oh, then I will call there, if you the telephone number give.
> oh, then I will call there if you give me the telephone number.

  I would like a winter vacation in Val-di-Fiemme plan
  for 2 people.
> I would like to plan a winter vacation in Val-di-Fiemme
> for 2 people.

  I can from my vacation place easy reach, right?
> can I reach from my vacation place easily, right?

  and can you say a hotel in case that
  could not possible for me?
> and can you tell me a hotel in case that apartment
> is not possible?
```

Experimental results

- improvements in word alignment quality
- translation results
- Verbmobil and Nespole! German-English tasks

Evaluation Methodology

- **word alignment quality: Alignment Error Rate (AER)**
  - **compare produced alignment connections A with reference alignment connections**
  - **Sure (S) and Possible (P) reference alignment connections exist, $S \subseteq P$**
  - **recall error: sure alignment is not found; precision error: a found alignment is not even possible**

$$\textbf{recall} = \frac{|A \cap S|}{|S|} \qquad \textbf{precision} = \frac{|A \cap P|}{|A|}$$

$$\textbf{AER}(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

- **translation results: automatic evaluation**
  - **Word Error Rate (WER)**
  - **Position-Independent Word Error Rate (PER)**
  - **BLEU score**

Verbmobil Alignment Training Corpus Statistics

- **Verbmobil German-English task, spontaneous speech**
- **domain: appointment scheduling, travel planning, hotel reservation**

		German	English
Train	Sentences	34K	
	Words	329 625	343 076
	Vocabulary	5 936	3 505
	Singletons	2 600	1 305
Dictionary	Entries	4 404	
Alignment test corpus	Sentences	354	
	Words	3 233	3 109

Results Verbmobil Task: smoothed lexicon

	German→English			English→German		
	Pre.[%]	Rec.[%]	AER [%]	Pre.[%]	Rec.[%]	AER [%]
34k Base	93.5	95.3	5.7	91.4	88.7	9.9
smooth	94.8	94.8	5.2	93.4	88.2	9.1
8k Base	92.5	95.4	6.2	88.7	88.3	11.5
smooth	93.2	94.9	6.0	89.9	87.8	11.1

- SMT system trained either on 34K or on 8K bilingual sentence pairs
- Method works better with larger training corpora
  (distribution of base forms can be better estimated)

Results Verbmobil Task: hierarchical lexicon counts

AER [%] corpus size = 0.5k				
Training	Model	$G \rightarrow E$	$E \rightarrow G$	combined
1^4H^5	hmm	18.8	24.0	16.9
	+hier	16.9	21.5	14.8
$1^4H^53^34^3$	ibm4	16.9	21.5	16.2
	+hier	15.8	20.7	14.9
$1^4H^53^34^36^5$	model6	16.7	21.1	15.9
	+hier	15.6	20.9	14.8

AER [%] corpus size = 34k				
Training	Model	$G \rightarrow E$	$E \rightarrow G$	combined
1^4H^5	hmm	8.9	14.9	7.9
	+hier	8.4	13.7	7.3
$1^4H^53^34^3$	ibm4	6.3	10.9	6.0
	+hier	6.1	10.8	5.7
$1^4H^53^34^36^5$	model6	5.7	9.9	5.5
	+hier	5.5	9.7	5.0

- **method is effective for small and large training corpora**
- **improvements are more significant for simpler alignment models**

Nespole! corpus statistics

- translation experiments on the Nespole! corpus of manually transcribed telephone inquiries (kindly provided by IRST)
- domain: travel information, hotel reservation
- training corpus extended with relevant in-domain data automatically selected from larger corpora
- n-gram coverage scores were used to select additional data

	German	English
Sentence pairs	3046	
Running words	14437	14743
Vocabulary	1452	1118
Singletons	734	472
Extension through n-gram coverage		
Sentence pairs	15835	
Running words	201907	207515
Vocabulary	17361	12367
Singletons	10423	4583

Translation results Nespole! Task

- compound splitting of German nouns performed in training and in testing
- test corpora statistics:

	Development	Test
Sentence pairs	300	106
Running words	1437	933
OOV-Rate	0.84 %	0.96 %

- results:

	WER [%]	PER [%]	BLEU
Baseline	60.7	47.4	0.212
+ in-domain corpus	56.1	45.2	0.238
+ sentence reordering (German)	53.7	45.5	0.270

- most improvements are in translation fluency

Translation Results Verbmobil Task

- training performed using the 8K training corpus
  to intensify the data sparseness problem
- test corpora statistics:

	Development	Test
Sentence pairs	276	251
Running words	3159	2628
OOV-Rate	3.3 %	4.0 %

- translation results:

	WER [%]	PER [%]	BLEU
Baseline	56.3	38.2	0.241
+ reordering (German)	52.3	37.9	0.261

Conclusions

Translation of speech with limited amount of training data:

- **a consistent way of selecting additional in-domain data from foreign sources**
- **two effective methods for inclusion of morpho-syntactic information in word alignment training to improve vocabulary coverage**
  - **morpho-syntactic information helped to improve alignment quality**
- **utilization of part-of-speech information to monotonize the translation process**
  - **significant improvements in translation fluency achieved on two tasks with highly spontaneous utterances**

Outlook

- goal: integrate the POS-based reordering in the search process
- perform experiments on automatically transcribed speech
- use syntax and morphology to reduce the Out-Of-Vocabulary rates