Simple and Effective Approach for Consistent Training of Hierarchical Phrase-based Translation Models

Stephan Peitz, David Vilar and Hermann Ney

`peitz@cs.rwth-aachen.de`

EACL 2014, Gotenburg
April 29th, 2014

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Motivation

- Statistical machine translation (SMT)
  - Language model
  - Translation model
    (set of bilingual phrases with translation probabilities)
- Extraction of bilingual phrases [Koehn & Och^{+} 03]
  - Given: word alignment, bilingual training data
  - Extract and count all valid phrases
  - Compute translation probabilities as relative frequencies
- Issues of this heuristic
  - Phrase extracted from likely alignment?
  - Models used in decoding are not considered $\Rightarrow$ inconsistency
- Solution for phrase-based SMT [Wuebker & Mauser^{+} 10]
- In this talk: consistent training for hierarchical phrase-based SMT

Hierarchical Phrase-based Translation

Translation model [Chiang 05]

- **Allow discontinuous phrases with "gaps"**
- **Obtain phrases from word-aligned bilingual training data**
  - **Sub-phrases within a phrase are replaced by a generic non-terminal X**
  - **Maximum of two gaps per rule**

$$X \rightarrow \langle \text{über } X_0 \text{ hinausgehen } X_1, \text{go beyond } X_0 \; X_1 \rangle$$

- **Reordering is modelled implicitly**
- **Formalized as a synchronous context-free grammar (SCFG)**
- **Speaking of *rules* rather than phrases**

Hierarchical Phrase-based Translation

Decoding [Chiang 07]

- **Parse input sentence with CYK+ algorithm [Chappelier & Rajman 98]**
  - **Use the source language part of the SCFG**
- **Get hypergraph representing all possible *derivations***
  - **Derivation: set of applied rules to generate an input sentence**
  - **Using the associated target part, translation can be constructed**
- **Incorporate language model (cube pruning algorithm)**

Consistent Translation Model Training

Main idea: Apply decoder on the training data

- Starting point: heuristically extracted translation model
- Run MERT [Och 03] on a development set to produce the baseline system
- Perform forced decoding on the training data
  - Translate source sentence to produce corresponding target sentence
- Extract k-best derivations and the rules applied in each derivation
- Recompute translation probabilities

Forced Decoding

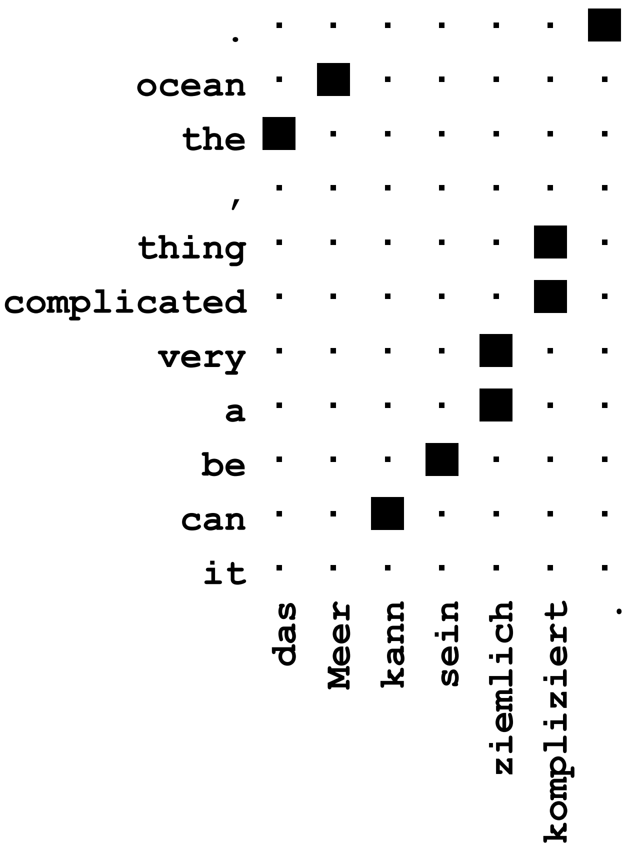
- Given a sentence pair (f_n, e_n) of the training data
- Constrain the translation of f_n
  - ***Force*** **the decoder** to produce e_n
- Simplification
  - Language model score is constant, incorporation is not needed
  - Cube pruning algorithm is unnecessary
  - **Forced decoding equals bilingual parsing** of the training data
- Less average run-time [Dyer 10]
  - Splitting one bilingual parse into two successive monolingual parses
  - First parse f_n, then the e_n

Forced Decoding

- (f_n, e_n) **has been parsed successfully**
- **Employy top-down k-best parsing algorithm [Chiang & Huang 05]**
  - **Find the k-best derivations**
  - **All models of the translation process are included (except for the language model)**
  - **Employ leave-one-out to counteract overfitting [Wuebker & Mauser$^+$ 10]**
- **Extract applied rules from k-best derivations**
  - **Count such rules**
  - **Recompute translation probabilities**

Example

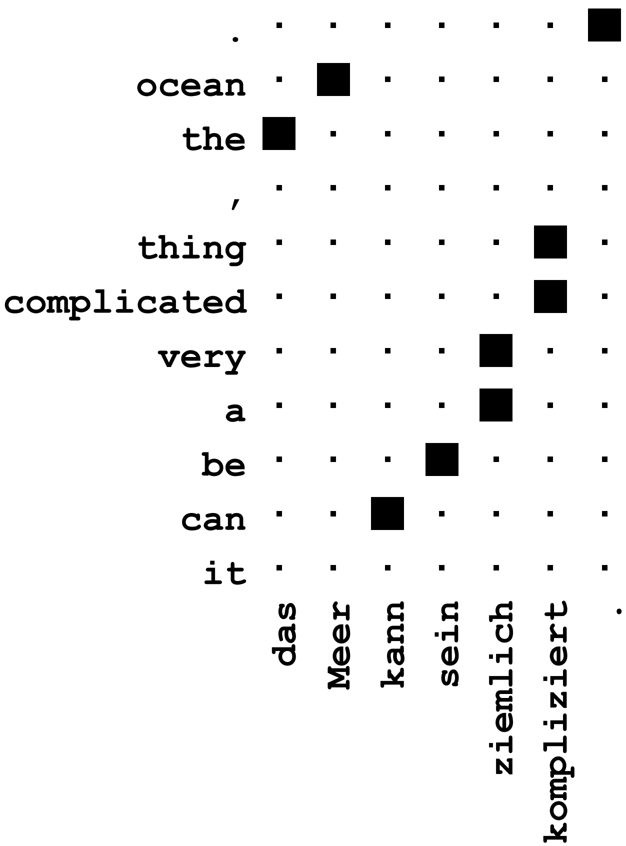
- **Heuristic extraction**




```
1 # das # , the
1 # das # the
1 # das Meer # , the ocean
1 # das Meer # the ocean
1 # das Meer kann sein X~0 # can be X~0 , the ocean
1 # das Meer kann sein X~0 # can be X~0 the ocean
1 # das Meer kann X~0 # it can X~0 , the ocean
1 # das Meer kann X~0 # it can X~0 the ocean
1 # das Meer kann X~0 # can X~0 , the ocean
1 # das Meer kann X~0 # can X~0 the ocean
2 # das Meer X~0 # X~0 , the ocean
2 # das Meer X~0 # X~0 the ocean
...
```

Example

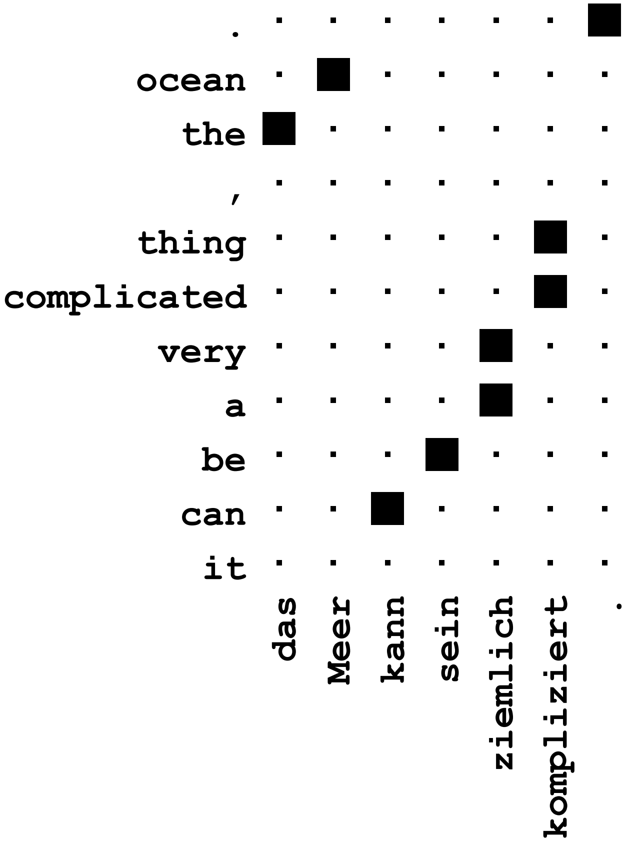
- **1-best forced derivation**




```
1 # . # .
1 # das # , the
1 # kann # it can
1 # ziemlich kompliziert # a very complicated thing
1 # X~0 sein X~1 # X~0 be X~1
1 # X~0 Meer X~1 # X~1 X~0 ocean
```

Example

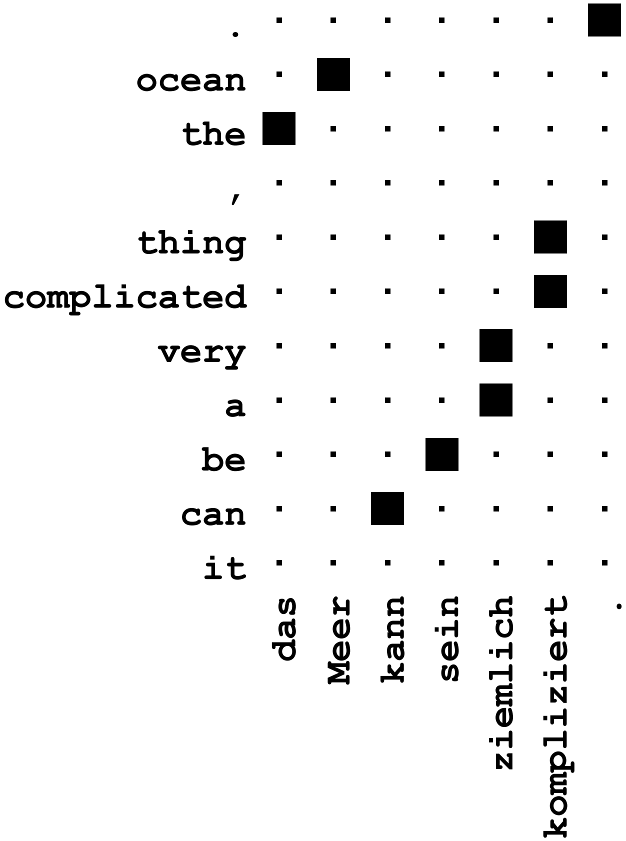
- **2-best forced derivations**



2 # . # .
2 # das # , the
1 # kann # it can
2 # ziemlich kompliziert # a very complicated thing
1 # X~0 sein X~1 # X~0 be X~1
2 # X~0 Meer X~1 # X~1 X~0 ocean
1 # kann sein X~0 # it can be X~0

Example

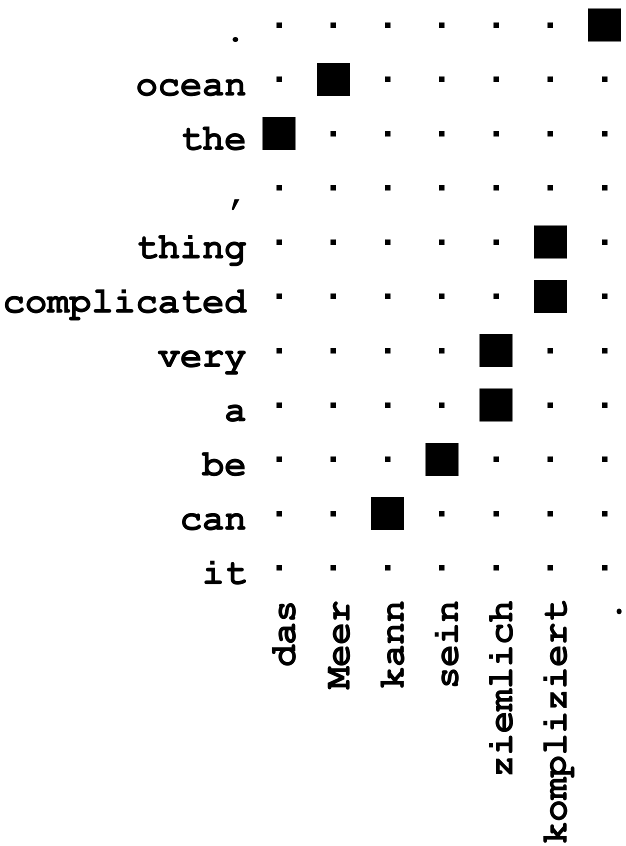
- 3-best forced derivations




```
3 # . # .
3 # das # , the
1 # kann # it can
3 # ziemlich kompliziert # a very complicated thing
1 # X~0 sein X~1 # X~0 be X~1
3 # X~0 Meer X~1 # X~1 X~0 ocean
1 # kann sein X~0 # it can be X~0
1 # kann X~0 # it can X~0
1 # sein X~0 # be X~0
```

Example

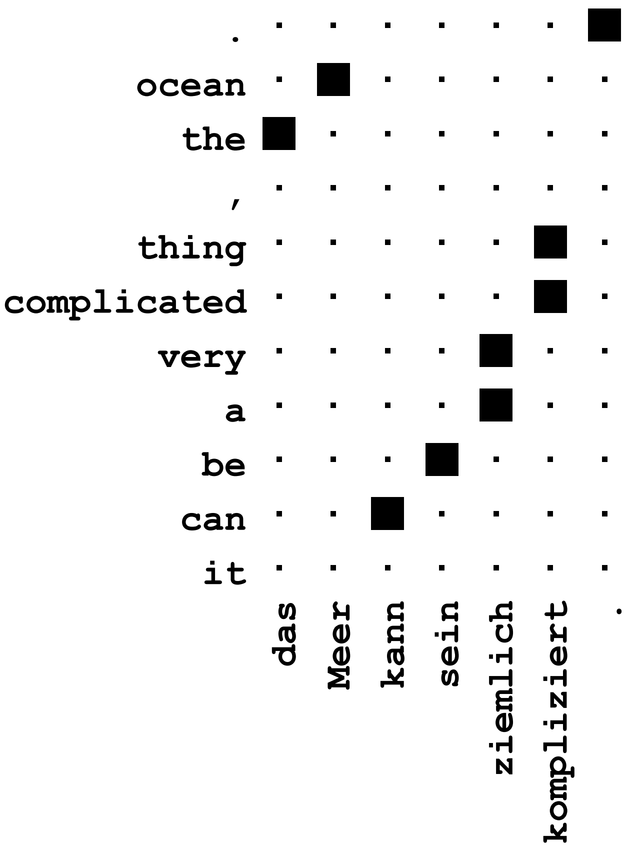
- **4-best forced derivations**




```
4 # . # .
4 # das # , the
1 # kann # it can
4 # ziemlich kompliziert # a very complicated thing
1 # X~0 sein X~1 # X~0 be X~1
4 # X~0 Meer X~1 # X~1 X~0 ocean
1 # kann sein X~0 # it can be X~0
1 # kann X~0 # it can X~0
1 # sein X~0 # be X~0
1 # kann sein # it can be
```

Example

- **5-best forced derivations**




```
5 # . # .
5 # das # , the
1 # kann # it can
4 # ziemlich kompliziert # a very complicated thing
1 # X~0 sein X~1 # X~0 be X~1
5 # X~0 Meer X~1 # X~1 X~0 ocean
1 # kann sein X~0 # it can be X~0
1 # kann X~0 # it can X~0
1 # sein X~0 # be X~0
1 # kann sein # it can be
1 # kann # it can be
1 # kompliziert # complicated thing
1 # X~0 sein X~1 # X~0 X~1
1 # X~0 ziemlich X~1 # X~0 a very X~1
```

Experiments

Setup

- IWSLT 2013 German→English
  - Translation of TED talks
  - 4.32M parallel sentences
  - 1.7 billion running words for LM training
- Forced decoding on indomain data
  - TED Talks
  - 140K sentences
  - Around 5% of the sentences could not be parsed

Experiments

Results

- Three independent runs of MERT [Och 03]
- Optimized on dev, BLEU as optimization criterion
- For forced decoding: $k = 500$

	dev		test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Hiero baseline	33.1	46.8	30.5	49.7
+ forced decoding	33.6	46.2	31.8	48.3

- Statistically significant improvements with at least 99% confidence
- Evaluated with *MultEval* [Clark & Dyer^{+} 11]

Conclusion

- A simple approach for consistent training of hierarchical translation models
  - Bilingual parsing
  - Top-down k-best derivation extraction
  - Recomputation of translation probabilities
- Effective: significant improvements of up to 1.3 points in BLEU
- Future work
  - Increase coverage

Implementation

```
            ('-.         .-') _   ('-.
           ( OO ).-.    ( OO ) )_(  OO)
     ,--.  / . --. /,--./ ,--,'(,------.     .-----.     .-----.
 .-')| ,|  | \-.  \ |   \ |  |\ |  .---'    / ,-.   \   /  -.   \
( OO |(_|.-'-'  |  ||    \|  | )|  |        '-'  |  |   '-' _'  |
| `-'|  | \| |_.'  ||  .     |/(|  '--.        .'  /       |_  <
,--. |  |  |  .-.  ||  |\    |  |  .--'      .'  /__    .-.  |  |
|  '-'  /  |  | |  ||  | \   |  |  `---.    |       |.-.\ `-'   /
 `-----'   `--' `--'`--'  `--'  `------'    `-------'`-' `----''
```

- RWTH's open-source translation toolkit
- new version Jane 2.3 includes
  - hierarchical decoder [Vilar & Stein+ 12]
  - phrase-based decoder [Wuebker & Huck+ 12]
  - system combination [Freitag & Huck+ 14]
  - forced alignment and forced derivation
- `http://www.hltpr.rwth-aachen.de/jane`

Thank you for your attention

Stephan Peitz

`peitz@cs.rwth-aachen.de`

`http://www-i6.informatik.rwth-aachen.de/~peitz`

References

[Chappelier & Rajman 98] J.C. Chappelier, M. Rajman: A generalized CYK algorithm for parsing stochastic CFG. In *Proceedings of the First Workshop on Tabulation in Parsing and Deduction*, pp. 133–137, April 1998. 4

[Chiang 05] D. Chiang: A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 263–270, Ann Arbor, Michigan, June 2005. 3

[Chiang 07] D. Chiang: Hierarchical phrase-based translation. *Computational Linguistics*, Vol. 33, No. 2, pp. 201–228, June 2007. 4

[Chiang & Huang 05] D. Chiang, L. Huang: Better k-best Parsing. In *Proceedings of the 9th Internation Workshop on Parsing Technologies*, pp. 53–64, Oct. 2005. 7

[Clark & Dyer^{+} 11] J.H. Clark, C. Dyer, A. Lavie, N.A. Smith: Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *ACL (Short Papers)*, pp. 176–181. The Association for Computer Linguistics, 2011. 15

[Dyer 10] C. Dyer: Two monolingual parses are better than one (synchronous parse). In *In Proc. of HLT-NAACL*, 2010. 6

[Freitag & Huck+ 14] M. Freitag, M. Huck, H. Ney: Jane: Open Source Machine Translation System Combination. In *Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April 2014. To appear. 17

[Huang & Zhou 09] S. Huang, B. Zhou: An EM algorithm for SCFG in formal syntax-based translation. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4813–4816, april 2009.

[Koehn & Och+ 03] P. Koehn, F.J. Och, D. Marcu: Statistical Phrase-Based Translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pp. 127–133, Edmonton, Alberta, 2003. 2

[Och 03] F.J. Och: Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, Sapporo, Japan, July 2003. 5, 15

[Vilar & Stein+ 12] D. Vilar, D. Stein, M. Huck, H. Ney: Jane: an advanced freely available hierarchical machine translation toolkit. *Machine Translation*, Vol. 26, No. 3, pp. 197–216, Sept. 2012. 17

[Wuebker & Huck+ 12] J. Wuebker, M. Huck, S. Peitz, M. Nuhn, M. Freitag, J.T. Peter, S. Mansour, H. Ney: Jane 2: Open Source Phrase-based and Hierarchical Statistical Machine Translation. In *International Conference on Computational Linguistics*, Mumbai, India, Dec. 2012. To appear. 17

[Wuebker & Mauser+ 10] J. Wuebker, A. Mauser, H. Ney: Training Phrase Translation Models with Leaving-One-Out. In *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, pp. 475–484, Uppsala, Sweden, July 2010. 2, 7

Experiments

Additional Results

	dev		eval		test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]	BLEU [%]	TER [%]
Hiero baseline	33.1	46.8	35.7	44.1	30.5	49.7
+ forced decoding -l1o	33.2	46.3	36.3	43.4	31.2	48.8
+ forced decoding +l1o	33.6	46.2	36.6	43.0	31.8	48.3
+ indomain TM	33.3	46.5	35.9	43.8	31.1	48.8

- Statistically significant improvements with at least 99% confidence