TECHNICAL UNIVERSITY OF VALENCIA (UPV)
DEPARTMENT OF COMPUTER SYSTEMS AND COMPUTATION (DSIC)

COMBINING TRANSLATION MODELS IN STATISTICAL MACHINE TRANSLATION

JESÚS ANDRÉS-FERRER jandres@dsic.upv.es
ISMAEL GARCÍA-VAREA ivarea@info-ab.uclm.es
FRANCISCO CASACUBERTA fcn@dsic.upv.es

Contents

1 Introduction

- Translate a source sentence $\mathbf{f} \in \mathbf{F}^*$ into a target sentence $\mathbf{e} \in \mathbf{E}^*$
- Brown et al. (1993) approached the problem of MT from a purely statistical point of view
- Pattern recognition problem with a set of classes $\mathbf{E}^*$
- Optimal Bayes' classification rule:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e} \in \mathbf{E}^*} \{p(\mathbf{e}|\mathbf{f})\} \tag{1}$$

- Applying Bayes' theorem $\Longrightarrow$ *inverse translation rule* (ITR):

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e} \in \mathbf{E}^*} \{p(\mathbf{e}) \cdot p(\mathbf{f}|\mathbf{e})\} \tag{2}$$

- The model problem
- The search problem: NP-hard (Knight, 1999; Udupa and Maji, 2006)
- Several search algorithms have been proposed to solve this problem efficiently (Brown and others, Wang and Waibel, 1997; Yaser and others, 1999; German and others, 2001; Jelinek, 1969; García-Varea and Casacuberta, 2001; Tillmann and Ney, 2003).

Introduction

- Many SMT systems (Och et al., 1999; Och and Ney, 2004; Koehn et al., 2003; Zens et al., 2002) have proposed the use of the *direct translation rule* (DTR):

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e} \in \mathbf{E}^*} \{p(\mathbf{e}) \cdot p(\mathbf{e}|\mathbf{f})\} \tag{3}$$

  - Heuristic version of the ITR
  - Easier search algorithm for some of the translation models
  - Its statistical theoretical foundation has not been clear for long time
  - (Andrés-Ferrer et al., 2007) have provided an explanation of its use within decision theory

2 Decision Theory

- A classification problem is a decision problem:
  - A set of *objects:* $\mathcal{X}$
  - A set of *classes or actions:* $\Omega = \{\omega_1, \ldots, \omega_C\}$ for each object x
  - A *loss function:* $\mathrm{l}(\omega_k|\mathbf{x}, \omega_j)$
- A classification system is *Classification function*: $\mathrm{c} : \mathcal{X} \rightarrow \Omega$
- The *conditional risk given* x*:*

$$R(\omega_k|\mathbf{x}) = \sum_{\omega_j \in \Omega} \mathrm{l}(\omega_k|\mathbf{x}, \omega_j)\, \mathrm{p}(\omega_j|\mathbf{x}) \tag{4}$$

- *Global risk* for a classification function:

$$R(\mathrm{c}) = \mathrm{E}_{\mathbf{x}}[R(\mathrm{c}(\mathbf{x})|\mathbf{x})] = \int_{\mathcal{X}} R(\mathrm{c}(\mathbf{x})|\mathbf{x})\, \mathrm{p}(\mathbf{x}) d\mathbf{x} \tag{5}$$

- Best system?
  - Minimise the global risk
  - Minimise the conditional risk for each $x \Longrightarrow$ minimise the global risk
  - *Bayes' classification rule :*

$$\hat{\mathrm{c}}(\mathbf{x}) = \arg\min_{\omega \in \Omega} R(\omega|\mathbf{x}) \tag{6}$$

  - For each loss function there is one optimal classification rule

2.1 Loss Function

❍ *Quadratic loss functions:*

$$\mathrm{l}(\omega_k|\mathbf{x},\omega_j) = \begin{cases} 0 & \omega_k = \omega_j \\ \epsilon(\mathbf{x},\omega_k,\omega_j) & \text{otherwise} \end{cases} \tag{7}$$

❑ Optimal classification rule:

$$\hat{\mathrm{c}}(\mathbf{x}) = \arg\min_{\omega_k \in \Omega} \sum_{\omega_j \neq \omega_k} \epsilon(\mathbf{x},\omega_k,\omega_j)\,\mathrm{p}(\omega_j|\mathbf{x}) \tag{8}$$

❑ Search space: $O(|\Omega|^2)$

❑ Can be prohibitive for some problems

➢ Rough approximations of the sum: $\sum_{\omega_j \neq \omega_k}$

➢ N-best lists

Loss Function

- *Linear loss functions:*

$$\mathrm{l}(\omega_k|\mathbf{x},\omega_j) = \begin{cases} 0 & \omega_k = \omega_j \\ \epsilon(\mathbf{x},\omega_j) & \text{otherwise} \end{cases} \tag{9}$$

  - $\epsilon(\cdot)$:
    - Depends on the object $\mathbf{x}$
    - Depends on the correct class ω_j
    - Does NOT *depend on the class proposed by the system* ω_k
  - Optimal classification Rule (Andrés-Ferrer et al., 2007):

$$\hat{\mathrm{c}}(\mathbf{x}) = \arg\max_{\omega\in\Omega} \{\mathrm{p}(\omega|\mathbf{x})\,\epsilon(\mathbf{x},\omega)\} \tag{10}$$

  - Search space: $O(|\Omega|)$

- The 0-1 *loss function* is usually assumed:

$$\mathrm{l}(\omega_k|\mathbf{x},\omega_j) = \begin{cases} 0 & \omega_k = \omega_j \\ 1 & \text{otherwise} \end{cases} \tag{11}$$

  - Optimal classification rule:

$$\hat{\mathrm{c}}(\mathbf{x}) = \arg\max_{\omega\in\Omega} \{\mathrm{p}(\omega|\mathbf{x})\} \tag{12}$$

  - Different kind of errors are not distinguished
  - Not specially appropriate in some cases:
    - Infinite class problems

3 Statistical Machine Translation

- SMT is a decision problem where:
  - Objects: $\mathbf{X} = \mathbf{F}^*$
  - Classes: $\Omega = \mathbf{E}^*$
  - Loss function: $\mathrm{l}(\mathbf{e}_k|\mathbf{f}, \mathbf{e}_j)$

- A 0-1 loss function is often assumed

- Classification rule for the 0-1 loss function:

$$\hat{\mathbf{e}} = \hat{\mathrm{c}}(\mathbf{f}) = \arg\max_{\mathbf{e}_k \in \Omega} \{\mathrm{p}(\mathbf{e}_k|\mathbf{f})\}$$

- Classification rule for the 0-1 loss function + Bayes' Theorem

$$\hat{\mathbf{e}} = \hat{\mathrm{c}}(\mathbf{f}) = \arg\max_{\mathbf{e}_k \in \Omega} \{\mathrm{p}(\mathbf{f}|\mathbf{e}_k)\,\mathrm{p}(\mathbf{e}_k)\}$$

- This loss function is not specially appropriate for SMT

- The set of classes is infinite enumerable

3.1 Quadratic Loss Function

- Quadratic loss function in STM:

$$\mathrm{l}(\mathbf{e}_k|\mathbf{f},\mathbf{e}_j) = \begin{cases} 0 & \mathbf{e}_k = \mathbf{e}_j \\ \epsilon(\mathbf{f},\mathbf{e}_k,\mathbf{e}_j) & \text{otherwise} \end{cases}$$

- Classification rule:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}_k \in \mathbf{E}^\star}{\arg\min} \sum_{\mathbf{e}_j \neq \mathbf{e}_k} \epsilon(\mathbf{f},\mathbf{e}_k,\mathbf{e}_j)\, \mathrm{p}(\mathbf{e}_j|\mathbf{f}) \tag{13}$$

- Allow to introduce the evaluation error metric:
  - $\mathrm{l}(\mathbf{e}_k|\mathbf{f},\mathbf{e}_j) = \mathrm{BLEU}(\mathbf{e}_k,\mathbf{e}_j)$
  - $\mathrm{l}(\mathbf{e}_k|\mathbf{f},\mathbf{e}_j) = \mathrm{WER}(\mathbf{e}_k,\mathbf{e}_j)$
- Metric loss functions (R. Schlüter and Ney, 2005)
- Quadratic search space
- Approximation: N-best lists (Kumar and Byrne, 2004)
- Introduce a kernel (Cortes et al., 2005) as the loss function

3.2 Linear Loss Functions

- Linear loss function:

$$\mathrm{l}(\mathbf{e}_k|\mathbf{f},\mathbf{e}_j) = \begin{cases} 0 & \mathbf{e}_k = \mathbf{e}_j \\ \epsilon(\mathbf{f},\mathbf{e}_j) & \text{otherwise} \end{cases}$$

- Classification rule:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}\in\mathbf{E}^\star} \{\mathrm{p}(\mathbf{e}|\mathbf{f})\,\epsilon(\mathbf{f},\mathbf{e})\}$$

- *Inverse translation rule (ITR):*
  - Using $\epsilon(\mathbf{f},\mathbf{e}_j) = 1$ and Bayes' theorem: $\Longrightarrow \hat{\mathbf{e}} = \arg\max_{\mathbf{e}_j\in\mathbf{E}^\star} \{\mathrm{p}(\mathbf{f}|\mathbf{e}_j)\,\mathrm{p}(\mathbf{e}_j)\}$
- *Direct translation rule (DTR):*
  - Using $\epsilon(\mathbf{f},\mathbf{e}_j) = p(\mathbf{e}_j) \Longrightarrow \hat{\mathbf{e}} = \arg\max_{\mathbf{e}_j\in\mathbf{E}^\star} \{\mathrm{p}(\mathbf{e}_j|\mathbf{f})\,\mathrm{p}(\mathbf{e}_j)\}$
- *Inverse form of DTR (IFDTR)*
  - Applying Bayes' theorem to DTR $\Longrightarrow \hat{\mathbf{e}} = \arg\max_{\mathbf{e}_j\in\mathbf{E}^*} \left\{\mathrm{p}(\mathbf{e}_j)^2\,\mathrm{p}(\mathbf{f}|\mathbf{e}_j)\right\}$
  - DTR and IFDTR a measure of model asymmetries
- *Direct and inverse translation rule (I&DTR):*
  - Using $\epsilon(\mathbf{f},\mathbf{e}_j) = p(\mathbf{f},\mathbf{e}_j) \Longrightarrow \hat{\mathbf{e}} = \arg\max_{\mathbf{e}_j\in\mathbf{E}^\star} \{\mathrm{p}(\mathbf{e}_j|\mathbf{f})\,\mathrm{p}(\mathbf{f}|\mathbf{e}_j)\,\mathrm{p}(\mathbf{e}_j)\}$

3.3 Log-Lineal Models

- Most of the current SMT systems use log-lineal models (Och and Ney, 2004; Marino et al., 2006):

$$p(\mathbf{e}|\mathbf{f}) \approx \frac{\exp\left[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{f},\mathbf{e})\right]}{\sum_{\mathbf{e}'} \exp\left[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{f},\mathbf{e}')\right]}$$

- Use the ITR with previous model to obtain the classification rule: $\hat{\mathbf{e}} = \arg\max_{\mathbf{e}\in\mathbf{E}^*} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{f},\mathbf{e})$
- Where h_m is usually the logarithmic of a statistical model that approximates a probability distribution ($h_m(\mathbf{f},\mathbf{e}) = \log \mathrm{p}_m(\mathbf{f}|\mathbf{e})$, $h_m(\mathbf{f},\mathbf{e}) = \log \mathrm{p}_m(\mathbf{e}|\mathbf{f})$, $h_m(\mathbf{f},\mathbf{e}) = \log \mathrm{p}_m(\mathbf{e})$, ...)
- Decision Theory also explains these models:
  - It can be understood as a linear loss function with:

$$\epsilon(\mathbf{f},\mathbf{e}) = p(\mathbf{e}\,|\,\mathbf{f})^{-1} \prod_{m=1}^{M} f_m(\mathbf{f},\mathbf{e})^{\lambda_i}$$

  - With $f_m(\mathbf{f},\mathbf{e}) = \exp[h_m(\mathbf{f},\mathbf{e})]$.
  - Define a family of functions depending on a hyperparameter($\boldsymbol{\lambda}_1^M$):

$$\left\{ p(\mathbf{e}\,|\,\mathbf{f})^{-1} \prod_{m=1}^{M} f_m(\mathbf{f},\mathbf{e})^{\lambda_i} \;\middle|\; \forall \lambda_i \;:\; i \in [1,M] \right\}$$

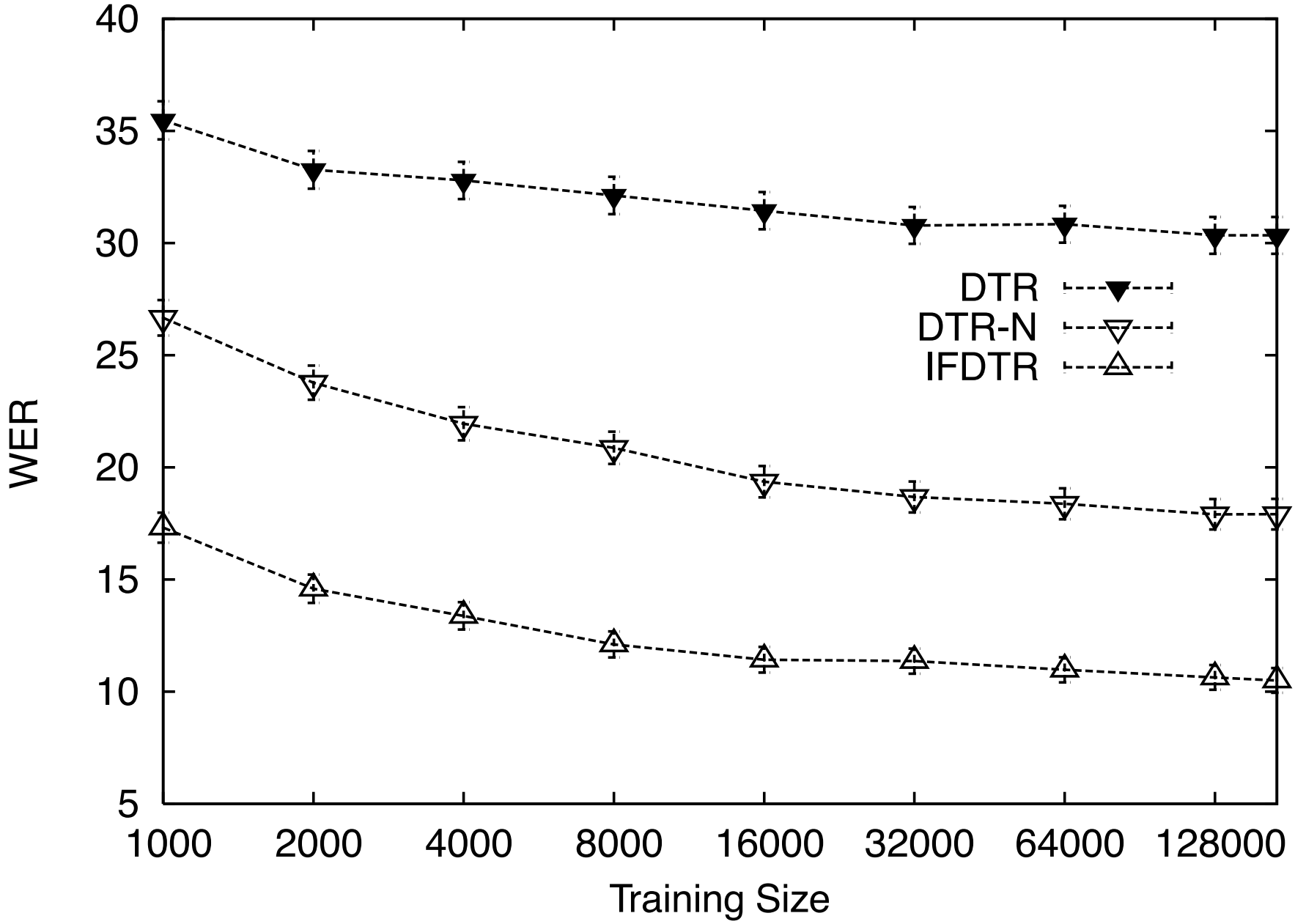
  - Experimentally (with a validation set) solve the optimisation problem
  - Use these hyperparameters to reduce the evaluation error metric (Och, 2003)

4 Experimental Results

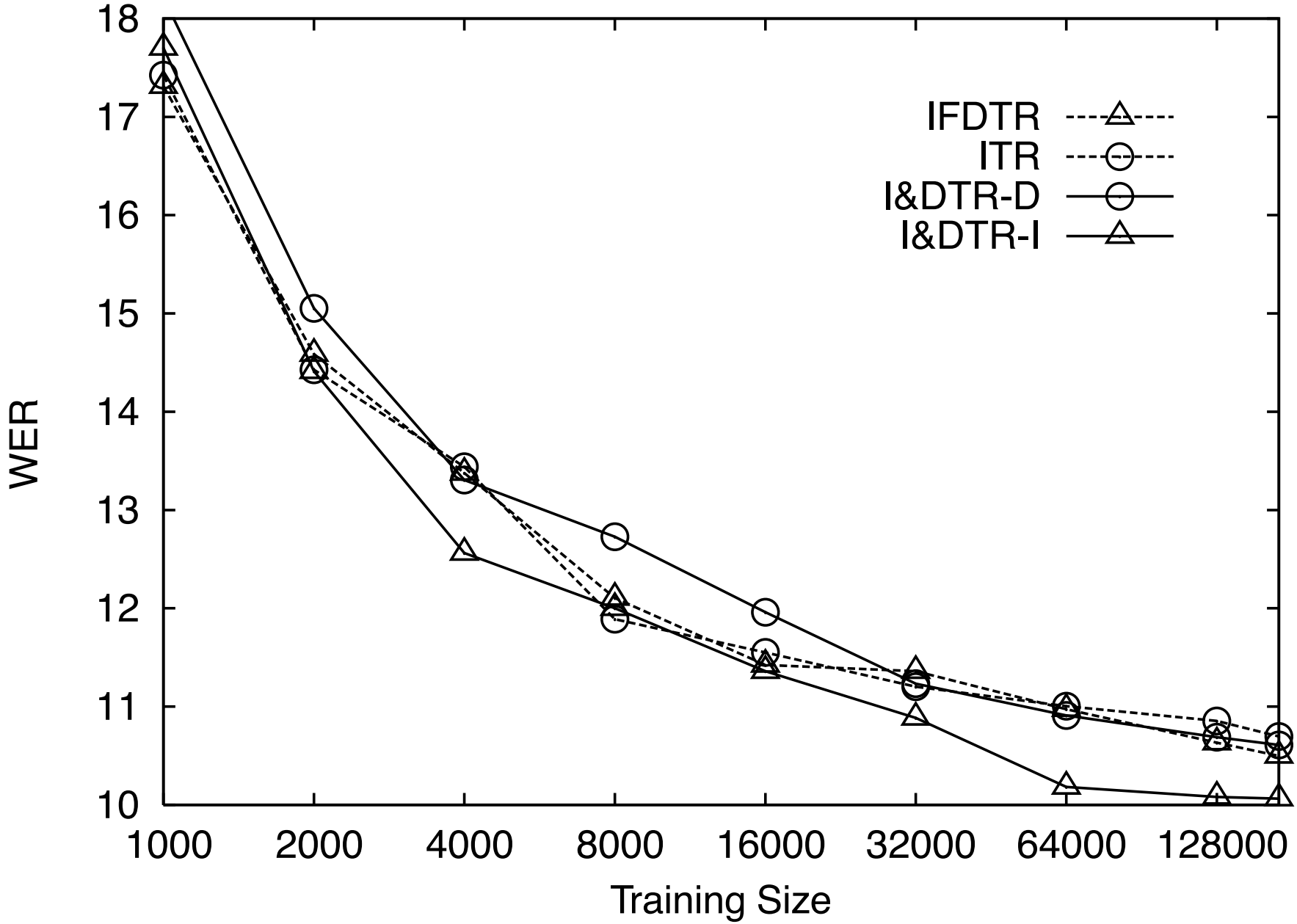
- Aim: Test theory in a small dataset and simple translation models
- State-of-art models in (Andrés-Ferrer et al., 2007)
- Results with IBM Model 2 (Brown and other, 1993) trained with *GIZA++* (Och, 2000)
- Decoding algorithm for each of the following rules (García-Varea and Casacuberta, 2001):
  - ITR: $\hat{\mathbf{e}} = \arg\max_{\mathbf{e}_j \in \mathbf{E}^\star} \{\mathrm{p}(\mathbf{f}|\mathbf{e}_j)\,\mathrm{p}(\mathbf{e}_j)\}$
  - DTR: $\hat{\mathbf{e}} = \arg\max_{\mathbf{e}_j \in \mathbf{E}^\star} \{\mathrm{p}(\mathbf{e}_j|\mathbf{f})\,\mathrm{p}(\mathbf{e}_j)\}$
  - IFDTR: $\hat{\mathbf{e}} = \arg\max_{\mathbf{e} \in \mathbf{E}^*} \{p(\mathbf{e})^2 p(\mathbf{f}|\mathbf{e})\}$
  - Two version of I&DTR (I&DTR-D and I&DTR-I): $\hat{\mathbf{e}} = \arg\max_{\mathbf{e}_j \in \mathbf{E}^\star} \{\mathrm{p}(\mathbf{e}_j|\mathbf{f})\,\mathrm{p}(\mathbf{f}|\mathbf{e}_j)\,\mathrm{p}(\mathbf{e}_j)\}$
- The Spanish-English TOURIST task (Amengual et al., 1996)
  - Human-to-human communication situations at the front-desk of a hotel
  - Semi-automatically produced using a small seed corpus from travel guides booklets
  - Test: 1K sentences randomly selected
  - Training sets of exponentially increasing sizes from 1K to 128K and 170K

	Test Set		Train Set	
	Spa	Eng	Spa	Eng
sentences	1K		170K	
avg. length	12.7	12.6	12.9	13.0
vocabulary	518	393	688	514
singletons	107	90	12	7
perplexity	3.62	2.95	3.50	2.89

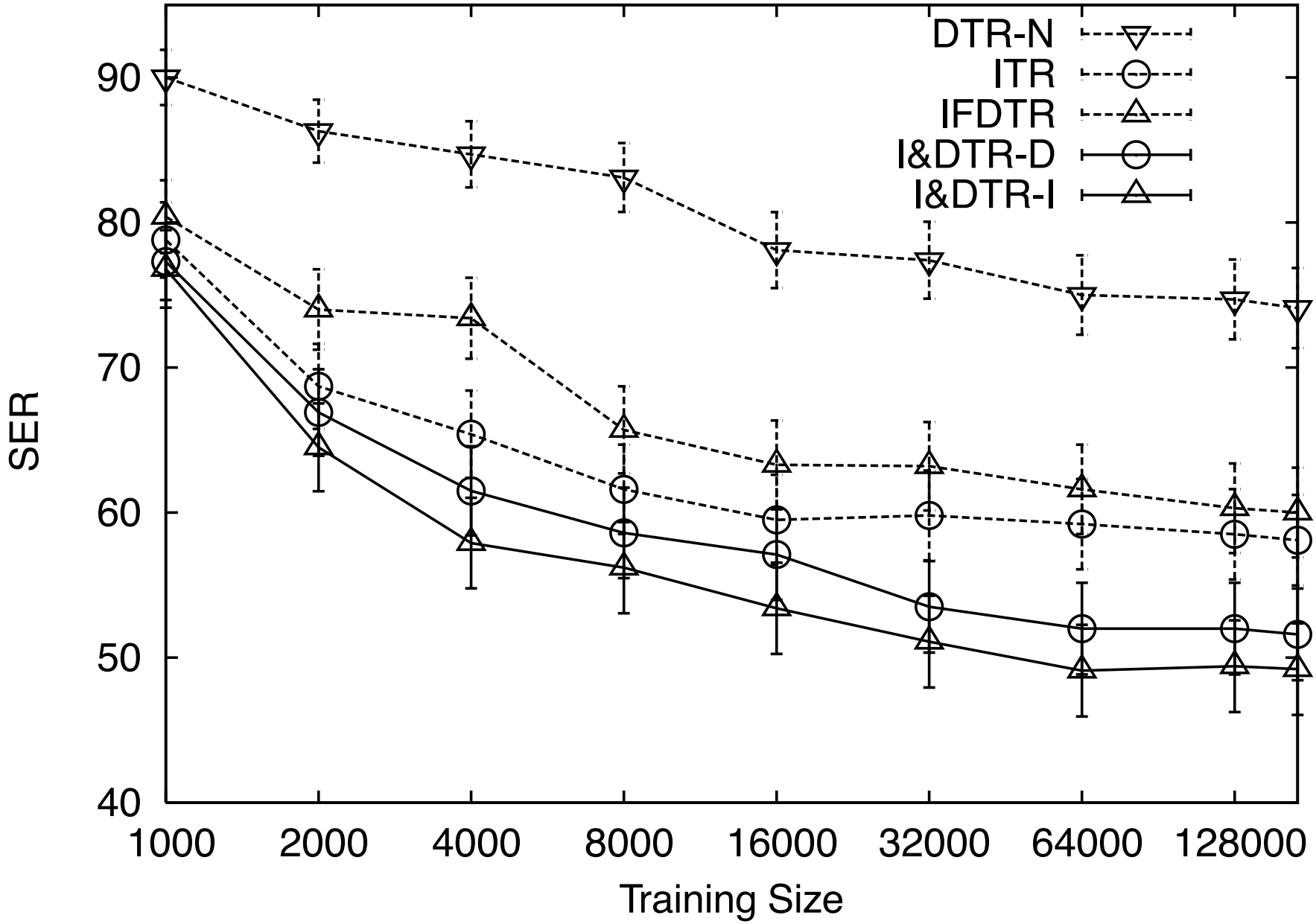
Asymmetry of Model 2



WER


WER
18
17
16
15
14
13
12
11
10
1000
2000
4000
8000
16000
32000
64000
128000
Training Size
IFDTR
ITR
I&DTR-D
I&DTR-I

SER


DTR-N
ITR
IFDTR
I&DTR-D
I&DTR-I
SER
90
80
70
60
50
40
1000
2000
4000
8000
16000
32000
64000
128000
Training Size

Global Results

- *Search error (SE)* (German and others, 2001): a translation error with a probability of the proposed translations less than the reference translation

- *Search error (ME):* a translation error with a probability of the proposed translations greater than the reference translation

Model	WER	SER	BLEU	SE	T
I&DTR I	**10.0**	**49.2**	**0.847**	1.3	34
I&DTR D	10.6	51.6	0.844	9.7	2
IFDTR	10.5	60.0	0.837	2.7	35
ITR	10.7	58.1	0.843	1.9	43
DTR N	17.9	74.1	0.750	0.0	2
DTR	30.3	92.4	0.535	0.0	2

5 Conclusions

- For each different loss function there is a different optimal Bayes' rule
- The most interesting loss functions incur in a Quadratic search space
- The classical 0-1 can be improved using a linear loss function
- The Framework explains the properties of some outstanding rules: ITR and DTR
- Some new rules have been proposed: I&DTR and IFDTR
- To increase performance, the best quadratic loss function should be found:

$$\mathrm{l}(\mathbf{e}_k|\mathbf{f},\mathbf{e}_j) = \begin{cases} 0 & \mathbf{e}_k = \mathbf{e}_j \\ \epsilon(\mathbf{f},\mathbf{e}_k,\mathbf{e}_j) & \text{otherwise} \end{cases} \tag{14}$$

- To increase performance keeping search space small, the best linear loss function should be found:

$$\mathrm{l}(\mathbf{e}_k|\mathbf{f},\mathbf{e}_j) = \begin{cases} 0 & \mathbf{e}_k = \mathbf{e}_j \\ \epsilon(\mathbf{f},\mathbf{e}_j) & \text{otherwise} \end{cases} \tag{15}$$

Thank you !

Questions ?

References

[Amengual et al.1996] J.C. Amengual, J.M. Benedí, M.A. Castaño, A. Marzal, F. Prat, E. Vidal, J.M. Vilar, C. Delogu, A. di Carlo, H. Ney, and S. Vogel. 1996. Definition of a machine translation task and generation of corpora. Technical report d4, Instituto Tecnológico de Informática, September. ESPRIT, EuTrans IT-LTR-OS-20268.

[Andrés-Ferrer et al.2007] J. Andrés-Ferrer, D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. 2007. On the use of different loss functions in statistical pattern recognition applied to machine translation. To appear in Pattern Recognition Letters.

[Brown and other1993] P. F. Brown and other. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

[Brown and others1990] P. F. Brown et al. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

[Cortes et al.2005] Corinna Cortes, Mehryar Mohri, and Jason Weston. 2005. A general regression technique for learning transductions. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 153–160, New York, NY, USA. ACM Press.

[García-Varea and Casacuberta2001] I. García-Varea and F. Casacuberta. 2001. Search algorithms for statistical machine translation based on dynamic programming and pruning techniques. In *Proc. of MT Summit VIII*, pages 115–120, Santiago de Compostela, Spain.

[German and others2001] U. German et al. 2001. Fast decoding and optimal decoding for machine translation. In *Proc. of ACL01*, pages 228–235.

[Jelinek1969] F. Jelinek. 1969. A fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development*, 13:675–685.

[Knight1999] Kevin Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615.

[Koehn et al.2003] P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edmonton, Canada, May.

[Kumar and Byrne2004] S. Kumar and W. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation.

[Marino et al.2006] J.B. Marino, R. E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. 2006. N-gram-based machine translation. In *Computational Linguistics*, pages 527–549.

[Och and Ney2004] F.J. Och and H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation . *Computational Linguistics*, 30(4):417–449, December.

[Och et al.1999] F. J. Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

[Och2000] F. J. Och. 2000. GIZA++: Training of statistical translation models. `http://www-i6.informatik.rwth-aachen.de/~och/software/GIZA++`

[Och2003] F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

[R. Schlüter and Ney2005] V. Steinbiss R. Schlüter, T. Scharrenbach and H. Ney. 2005. Bayes risk minimization using metric loss functions. In *Proceedings of the European Conference on Speech Communication and Technology, Interspeech*, pages 1449–1452, Lisbon, Portugal, September.

[Tillmann and Ney2003] Christoph Tillmann and Hermann Ney. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133, March.

[Udupa and Maji2006] Raghavendra Udupa and Hemanta K. Maji. 2006. Computational complexity of statistical machine translation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 25–32. Trento, Italy.

[Wang and Waibel1997] Ye-Yi Wang and Alex Waibel. 1997. Decoding algorithm in statistical translation. In *Proc. of ACL'97*, pages 366–372, Madrid, Spain.

[Yaser and others1999] A. Yaser et al. 1999. Statistical Machine Translation: Final Report. Technical report, Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, USA.

[Zens et al.2002] R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Advances in artificial intelligence. 25. Annual German Conference on AI*, volume 2479 of *Lecture Notes in Computer Science*, pages 18–32. Springer Verlag, September.