COMBINING TRANSLATION MODELS IN STATISTICAL MACHINE TRANSLATION

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

1 Introduction 0

2 Decision Theory 2
   2.1 Loss Function .......................................................... 3

3 Statistical Machine Translation 5
   3.1 Quadratic Loss Function .................................................. 6
   3.2 Linear Loss Functions .................................................... 7
   3.3 Log-Lineal Models .......................................................... 8

4 Experimental Results 9

5 Conclusions 14

Jesús Andrés Ferrer 1 DSIC ITI UPV TMI 2007
1 Introduction

- Translate a source sentence \( f \in F^* \) into a target sentence \( e \in E^* \)

- Brown et al. (1993) approached the problem of MT from a purely statistical point of view

- Pattern recognition problem with a set of classes \( E^* \)

- Optimal Bayes’ classification rule:

\[
\hat{e} = \arg \max_{e \in E^*} \{ p(e|f) \} 
\]

(1)

- Applying Bayes’ theorem \( \Rightarrow \) inverse translation rule (ITR):

\[
\hat{e} = \arg \max_{e \in E^*} \{ p(e) \cdot p(f|e) \}
\]

(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):

\[ \hat{e} = \arg \max_{e \in E^*} \{ p(e) \cdot p(e|f) \} \]  

- 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: $l(\omega_k|x, \omega_j)$

- A classification system is Classiﬁcation function: $c : \mathcal{X} \to \Omega$

- The conditional risk given $x$:

$$R(\omega_k|x) = \sum_{\omega_j \in \Omega} l(\omega_k|x, \omega_j) p(\omega_j|x)$$  \hspace{1cm} (4)

- Global risk for a classiﬁcation function:

$$R(c) = E_x[R(c(x)|x)] = \int_{\mathcal{X}} R(c(x)|x) p(x) dx$$  \hspace{1cm} (5)

- Best system?
  - Minimise the global risk
  - Minimise the conditional risk for each $x \Rightarrow$ minimise the global risk
  - Bayes’ classiﬁcation rule:

$$\hat{c}(x) = \arg \min_{\omega \in \Omega} R(\omega|x)$$  \hspace{1cm} (6)

  - For each loss function there is one optimal classiﬁcation rule
2.1 Loss Function

- **Quadratic loss functions:**

\[
l(\omega_k|x, \omega_j) = \begin{cases} 
0 & \omega_k = \omega_j \\
\epsilon(x, \omega_k, \omega_j) & \text{otherwise}
\end{cases}
\]  

(7)

- Optimal classification rule:

\[
\hat{c}(x) = \arg\min_{\omega_k \in \Omega} \sum_{\omega_j \neq \omega_k} \epsilon(x, \omega_k, \omega_j) p(\omega_j|x)
\]

(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
Linear loss functions:

\[ l(\omega_k|x, \omega_j) = \begin{cases} 
0 & \omega_k = \omega_j \\
\epsilon(x, \omega_j) & \text{otherwise}
\end{cases} \quad (9) \]

- \( \epsilon(\cdot) : \)
  - Depends on the object \( x \)
  - Depends on the correct class \( \omega_j \)
  - Does NOT depend on the class proposed by the system \( \omega_k \)

Optimal classification Rule (Andrés-Ferrer et al., 2007):

\[ \hat{c}(x) = \arg \max_{\omega \in \Omega} \left\{ p(\omega|x) \epsilon(x, \omega) \right\} \quad (10) \]

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

The 0-1 loss function is usually assumed:

\[ l(\omega_k|x, \omega_j) = \begin{cases} 
0 & \omega_k = \omega_j \\
1 & \text{otherwise}
\end{cases} \quad (11) \]

Optimal classification rule:

\[ \hat{c}(x) = \arg \max_{\omega \in \Omega} \left\{ p(\omega|x) \right\} \quad (12) \]

- 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: $X = F^*$
  - Classes: $\Omega = E^*$
  - Loss function: $l(e_k|f, e_j)$

- A 0-1 loss function is often assumed

- Classification rule for the 0-1 loss function:
  $$\hat{e} = \hat{c}(f) = \arg\max_{e_k \in \Omega} \{ p(e_k|f) \}$$

- Classification rule for the 0-1 loss function + Bayes’ Theorem
  $$\hat{e} = \hat{c}(f) = \arg\max_{e_k \in \Omega} \{ p(f|e_k) p(e_k) \}$$

- 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:

\[ l(e_k|f, e_j) = \begin{cases} 
0 & e_k = e_j \\
\epsilon(f, e_k, e_j) & \text{otherwise}
\end{cases} \]

- Classification rule:

\[ \hat{e} = \arg \min_{e_k \in E^*} \sum_{e_j \neq e_k} \epsilon(f, e_k, e_j) p(e_j|f) \] (13)

- Allow to introduce the evaluation error metric:
  - \( l(e_k|f, e_j) = \text{BLEU}(e_k, e_j) \)
  - \( l(e_k|f, e_j) = \text{WER}(e_k, e_j) \)

- 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:
  \[
  l(e_k|f, e_j) = \begin{cases} 
  0 & e_k = e_j \\
  \epsilon(f, e_j) & \text{otherwise}
  \end{cases}
  \]

- Classification rule:
  \[
  \hat{e} = \arg\max_{e \in E^*} \{p(e|f) \epsilon(f, e)\}
  \]

- Inverse translation rule (ITR):
  - Using \(\epsilon(f, e_j) = 1\) and Bayes’ theorem: \(\hat{e} = \arg\max_{e_j \in E^*} \{p(f|e_j) p(e_j)\}\)

- Direct translation rule (DTR):
  - Using \(\epsilon(f, e_j) = p(e_j)\) \(\hat{e} = \arg\max_{e_j \in E^*} \{p(e_j|f) p(e_j)\}\)

- Inverse form of DTR (IFDTR)
  - Applying Bayes’ theorem to DTR \(\hat{e} = \arg\max_{e_j \in E^*} \{p(e_j)^2 p(f|e_j)\}\)
  - DTR and IFDTR a measure of model asymmetries

- Direct and inverse translation rule (I&DTR):
  - Using \(\epsilon(f, e_j) = p(f, e_j)\) \(\hat{e} = \arg\max_{e_j \in E^*} \{p(e_j|f) p(f|e_j) p(e_j)\}\)
3.3 Log-Lineal Models

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

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

Use the ITR with previous model to obtain the classification rule:

$$\hat{e} = \arg \max_{e \in E^*} \sum_{m=1}^{M} \lambda_m h_m(f, e)$$

Where $h_m$ is usually the logarithmic of a statistical model that approximates a probability distribution ($h_m(f, e) = \log p_m(f|e)$, $h_m(f, e) = \log p_m(e|f)$, $h_m(f, e) = \log p_m(e)$, ...)

Decision Theory also explains these models:

- It can be understood as a linear loss function with:

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

- With $f_m(f, e) = \exp[h_m(f, e)]$.

- Define a family of functions depending on a hyperparameter($\lambda_1^M$):

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

- 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:** \( \hat{e} = \arg \max_{e_j \in E^*} \{p(f|e_j) p(e_j)\} \)
  - **DTR:** \( \hat{e} = \arg \max_{e_j \in E^*} \{p(e_j|f) p(e_j)\} \)
  - **IFDTR:** \( \hat{e} = \arg \max_{e \in E^*} \{p(e)^2 p(f|e)\} \)
  - **Two version of I&DTR (I&DTR-D and I&DTR-I):** \( \hat{e} = \arg \max_{e_j \in E^*} \{p(e_j|f) p(f|e_j) p(e_j)\} \)
- **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

<table>
<thead>
<tr>
<th></th>
<th>Test Set</th>
<th>Train Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spa</td>
<td>Eng</td>
</tr>
<tr>
<td>sentences</td>
<td>1K</td>
<td>170K</td>
</tr>
<tr>
<td>avg. length</td>
<td>12.7</td>
<td>12.6</td>
</tr>
<tr>
<td>vocabulary</td>
<td>518</td>
<td>393</td>
</tr>
<tr>
<td>singletons</td>
<td>107</td>
<td>90</td>
</tr>
<tr>
<td>perplexity</td>
<td>3.62</td>
<td>2.95</td>
</tr>
</tbody>
</table>
Asymmetry of Model 2

![Graph showing WER vs Training Size for different models: DTR, DTR-N, IFDTR. The graph indicates that DTR has the lowest WER across all training sizes.](image)
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

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
<th>SER</th>
<th>BLEU</th>
<th>SE</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>I&amp;DTR I</td>
<td>10.0</td>
<td>49.2</td>
<td>0.847</td>
<td>1.3</td>
<td>34</td>
</tr>
<tr>
<td>I&amp;DTR D</td>
<td>10.6</td>
<td>51.6</td>
<td>0.844</td>
<td>9.7</td>
<td>2</td>
</tr>
<tr>
<td>IFDTR</td>
<td>10.5</td>
<td>60.0</td>
<td>0.837</td>
<td>2.7</td>
<td>35</td>
</tr>
<tr>
<td>ITR</td>
<td>10.7</td>
<td>58.1</td>
<td>0.843</td>
<td>1.9</td>
<td>43</td>
</tr>
<tr>
<td>DTR N</td>
<td>17.9</td>
<td>74.1</td>
<td>0.750</td>
<td>0.0</td>
<td>2</td>
</tr>
<tr>
<td>DTR</td>
<td>30.3</td>
<td>92.4</td>
<td>0.535</td>
<td>0.0</td>
<td>2</td>
</tr>
</tbody>
</table>
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:

$$l(e_k|f, e_j) = \begin{cases} 
0 & e_k = e_j \\
\epsilon(f, e_k, e_j) & \text{otherwise}
\end{cases}$$  \hspace{1cm} (14)

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

$$l(e_k|f, e_j) = \begin{cases} 
0 & e_k = e_j \\
\epsilon(f, e_j) & \text{otherwise}
\end{cases}$$ \hspace{1cm} (15)
Thank you!
Questions ?
References


