Word Confidence Estimation for Machine Translation
Ngoc Quang Luong, Laurent Besacier, Benjamin Lecouteux

To cite this version:
Ngoc Quang Luong, Laurent Besacier, Benjamin Lecouteux. Word Confidence Estimation for Machine Translation. Journées du LIG, 2013, Grenoble, France. hal-02094767

HAL Id: hal-02094767
https://hal.archives-ouvertes.fr/hal-02094767
Submitted on 9 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
INTRODUCTION

Definition:
• Confidence Estimation (CE) is a task of judging automatically each part (e.g., word, segment, or the whole sentence) in the MT hypothesis as correct or incorrect.
• A classifier trained beforehand by a feature set calculates the confidence score for MT hypothesis, then compares it with a threshold. Those with scores exceeding this threshold are categorized in the Good label set; the rest belongs to the Bad label set.

Interesting uses of CE:
• Decide whether a given translation is good enough for publishing as is.
• Highlight words that need editing in post-editing tasks.
• Inform readers of portions of the sentence that are not reliable.
• Select the best segments among options from multiple translation systems for MT system combination.

FIRST EXPERIMENTS AND RESULTS

1. Preliminary experiment with all features:
   • We track the Precision (Pr), Recall (Rc) and F-score (F) values for G and B label along threshold variation (from 0.3 to 1.0, step 0.025).
   • Compare to 2 baselines: Baseline 1 (all words in each MT hypothesis are classified as good), and Baseline 2 (assigned randomly 85%G + 15%B) (Table 2, left, below).

<table>
<thead>
<tr>
<th>System</th>
<th>Label</th>
<th>Pr(%)</th>
<th>Rc(%)</th>
<th>F(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>Good</td>
<td>86.02</td>
<td>88.07</td>
<td>87.04</td>
</tr>
<tr>
<td></td>
<td>Bad</td>
<td>39.11</td>
<td>35.41</td>
<td>37.17</td>
</tr>
<tr>
<td>Baseline 1</td>
<td>Good</td>
<td>81.78</td>
<td>100.00</td>
<td>89.98</td>
</tr>
<tr>
<td></td>
<td>Bad</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline 2</td>
<td>Good</td>
<td>81.77</td>
<td>85.20</td>
<td>83.45</td>
</tr>
<tr>
<td></td>
<td>Bad</td>
<td>18.14</td>
<td>14.73</td>
<td>16.26</td>
</tr>
</tbody>
</table>

Table 3: Example of all-feature classifier’s output

2. Feature Selection:
   • Objective: to rank our features from most to least important + to find the best performing combination.
   • Strategy: “Sequential Backward Selection” algorithm. We start from the full set of N features, and in each step sequentially remove the most useless one.
   • Output: The rank of each feature (also its ID in Table 1) + the system’s evolution as more and more features are removed (Figure 1, above on the right).

3. Boosting technique to improve the classifier’s performance:
   • Objective: Take advantage of the sub-models’ complementarily when combined.
   • How to prepare the training set for Boosting system:
     – Step 1: Starting from 25 features, we build 23 subsets, in which 1 contains all features, 1 contains top 10 in Table 1, and 21 sets of 9 randomly extracted features for each.
     – Step 2: Divide our 10K training set into 10 equal subsets (S1, S2, ..., S10).
     – Step 3: For i=1 to 10 do
       – Concatenate Si into Si (i=1, 10, 11) Train this set by 23 feature above sets (sequentially) => 23 different classifiers.
     – Step 4: Concatenate Di into D1 into D2 to Di into Di to obtain the Boosting training set.
   • Testing: Build the test set for Boosting by logging 23 scores (like Step 3) for the usual test set, coming from 23 systems built on the usual training set.
   • Comparison of the performance between 2 systems in terms of averaged scores (Table 3, below) or scores along to threshold variation (Figure 2, right).

CONCLUSION AND ONGOING RESEARCH

• Experimental results show that precision and recall obtained in Good label are very promising, and Bad label reaches acceptable performance.
• A feature selection that we proposed helped to identify the most valuable features, as well as to find out the best performing subset among them.
• The protocol of applying Boosting method exploited effectively the good feature subsets for the system’s performance improvement.
• Future work will examine the linguistic features of word; experiment the system at segment level and find the methodology to conclude the sentence quality rely on the word’s and segment’s confidence score.