Quality Estimation for Synthetic Parallel Data Generation

Raphael Rubino*, Antonio Toral†, Nikola Ljubešić‡, Gema Ramírez-Sánchez*
* Prompsit Language Engineering, Elche, Spain
† CNGL – School of Computing, Dublin City University, Ireland
‡ Department of Information and Communication Sciences, University of Zagreb, Croatia
{rubino, gramirez}@prompsit.com, atoral@computing.dcu.ie, nljubes@ffzg.hr

Abstract

This paper presents a novel approach for parallel data generation using machine translation and quality estimation. Our study focuses on pivot-based machine translation from English to Croatian through Slovene. We generate an English–Croatian version of the Europarl parallel corpus based on the English–Slovene Europarl corpus and the Apertium rule-based translation system for Slovene–Croatian. These experiments are to be considered as a first step towards the generation of reliable synthetic parallel data for under-resourced languages. We first collect small amounts of aligned parallel data for the Slovene–Croatian language pair in order to build a quality estimation system for sentence-level Translation Error Rate (TER) estimation. We then infer TER scores on automatically translated Slovene to Croatian sentences and use the best translations to build an English–Croatian statistical MT system. We show significant improvement in terms of automatic metrics obtained on two test sets using our approach compared to a random selection of synthetic parallel data.

Keywords: Under-resourced Languages, Synthetic Corpora, Machine Translation, Quality Estimation

1. Introduction

Previous work on synthetic parallel data generation relies on the use of machine translation (MT) to translate source text into the target language for a given language pair in order to obtain a new parallel corpus. This resource can then be used as training material for SMT, or any other application that requires parallel data. However, one important limitation of this artificial resource is its translation quality. As it is directly related to the performance of data-driven systems, the need for estimating the translation quality of synthetically built corpora seems obvious.

This paper applies quality estimation (QE) techniques for the generation of synthetic parallel data. Our case study is on the English–Croatian language pair with the Slovene language as pivot. We first train a Slovene–Croatian QE system by collecting limited amounts of parallel data for these languages from diverse sources. Then, the source side of the corpus is translated using the Apertium rule-based MT (RBMT) system [Forcada et al., 2011]. The translated text is compared to its reference (the target side of the corpus) at the sentence level using TER [Snover et al., 2006]. With these scores as labels, a regression model is built on feature vectors representing the sentence pairs (source–translation). Using the regression model, TER scores are inferred on automatically translated Slovene to Croatian sentences taken from the English–Slovene Europarl parallel corpus [Koehn, 2005]. The best translations are used to build an English–Croatian statistical MT (SMT) system.

After giving an overview of previous work in the areas of pivot-based MT and QE in Section 2.1, the first part of our study is to build and evaluate a QE model for Slovene to Croatian, described in Section 3. We then present the SMT setup for translating from English to Croatian and the results obtained using synthetic data in Section 4. Finally, we conclude and give details about future work in Section 5.

2. Previous Work

2.1. Synthetic Data for Pivot-based MT

Pivot-based MT refers to the use of an intermediate language, called pivot language (PL), to translate from the source- (SL) to the target language (TL). Differently than typical MT systems, which translate directly from SL to TL, pivot-based systems translate sequentially from SL to PL and then from PL to TL. The main motivation for building pivot-based MT systems is the lack of language resources for a language pair SL–TL, in contrast with the availability of such resources for both language pairs SL–PL and PL–TL.

This is our case as our aim is to translate from English to Croatian, but to do so we use Slovene as a pivot. Our bilingual resources are for the English–Slovene language pair (Europarl parallel corpus) and for Slovene–Croatian (RBMT system).

Pivot-based strategies in MT can be classified into three categories [Wu and Wang, 2009]: phrase table multiplication (also known as triangulation), transfer (also referred to as cascade) and synthetic corpus.

The synthetic corpus approach [Gispert and Mariño, 2006; Bertoldi et al., 2008; Utiyama et al., 2008] is the one we work upon. In this method a SL–TL corpus is obtained using the SL–PL or the PL–TL corpora. One way to do this is to translate the PL sentences in the SL–PL corpus into TL with the PL–TL system. Another possibility is to translate the PL sentences in the PL–TL corpus into SL with the SL–PL system. Obviously, both methods could be applied and the two resulting synthetic corpora be merged into a single SL–TL corpus.

In this paper we extend the synthetic corpus approach to pivot-based MT by filtering the resulting synthetic corpus with QE.

2.2. Quality Estimation and Applications

Estimating the quality of MT output is the ability to judge the correctness of a translation without any translation ref-
ference. Since the first work conducted on QE for MT at the
word and sentence levels, this task has grown in interest and
performance in the past few years. (Gandrabur and Foster,
2003; Uetzing et al., 2003; Blatz et al., 2003) Recent eval-
uation campaigns helped defining QE baselines and state-
of-the-art systems, based on supervised learning using vec-
torial representations of source sentences and their trans-
lations associated with quality scores or labels. (Callison-
Burch et al., 2012; Bojar et al., 2013) The usefulness of fea-
tures is directly related to the QE task itself, i.e. varies
to generalise well and avoid training data overfitting, we
then extend this feature set in order to improve the QE per-
formance and measure its impact on the synthetic parallel
data selection.

3. Quality Estimation for Slovene–Croatian
The QE setup designed for our experiments on synthetic
parallel data generation is presented in this Section. We first
introduce the data and tools required to build and evaluate
the QE models in Subsection 3.1, followed by the feature
sets described in Subsection 3.2, extracted from the text
data. Finally, the QE models evaluation results are detailed
in Subsection 3.3.

3.1. Dataset and Tools
In order to build and evaluate QE models for the Slovene–
Croatian language pair, we collect three parallel corpora for
these languages:

- the EAC Translation Memory (noted EAC) containing
  573 translation units,
- the EU Bookshop parallel corpus (noted EUb) containing
  4,222 sentence pairs,
- a small Slovene–Croatian parallel corpus obtained from a translation agency (noted slhr) containing
  2,286 sentence pairs.

We first consider these corpora individually to build and
evaluate three QE models, before concatenating the data
(noted all) into one corpus and building our final QE model.
This way, four QE models are trained, and evaluated on four
test sets. We present the four corpora used for our QE ex-
periment in Table 1. For each of the parallel corpora, the
source sentences are translated from Slovene to Croatian
using the Apertium RBMT system for this language direc-
tion. Source sentences, their translations and references
are then tokenised and lowercased using the tools provided
with the Moses MT system (Koehn et al., 2007). The tool
TERCOM provides us with the sentence-level TER scores.
This scores can be used as labels to train and evaluate our
QE models. Finally, sentence triplets are randomised and
the corpus is split in two parts: a training and a test set.

<table>
<thead>
<tr>
<th></th>
<th>EAC</th>
<th>EUb</th>
<th>slhr</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>500</td>
<td>3,000</td>
<td>2,000</td>
<td>5,500</td>
</tr>
<tr>
<td>Test</td>
<td>73</td>
<td>1,222</td>
<td>285</td>
<td>1,580</td>
</tr>
</tbody>
</table>

Table 1: Number of sentences in each configuration for the
three different corpora used in our experiments. The col-
umn all is the concatenation of the three other corpora.

Based on source sentences, their translations and their cor-
responding sentence-level TER scores, we train regression
models that aim to predict sentence-level TER scores on
unseen data. However, using words directly, or n-grams,
as features for QE usually leads to large and sparse vectors
which complicates the supervised learning step. In order
to generalise well and avoid training data overfitting, we
extract a tried-and-tested set of features described in Sec-
Section 3.2, using an in-house feature extractor tool-kit. We
consider this first set of 15 features as our baseline. We
then extend this feature set in order to improve the QE per-
formance and measure its impact on the synthetic parallel
data selection.

Regression models are trained using the $\epsilon$-SVR implemen-
tation available in the LibSVM toolkit (Chang and Lin,
2011). SVM parameters, namely $c$, $\gamma$ and $\epsilon$, are optimised
based on a 5-fold cross-validation approach using the train-
ing set. The best parameter triplet is chosen according to
several metrics: Mean Average Error (MAE), Root Mean
Square Error (RMSE), Pearson’s correlation coefficient ($r$)
and the total number of support vectors. In our experi-
ments, minimising the MAE and RMSE is not as crucial as
maximising the Pearson’s correlation coefficient, as the aim
of our work is to predict TER scores following a similar dis-
tribution as the reference ones.

3.2. Quality Estimation Features
The features extracted from each sentence pair, i.e. source
sentences and their translations, are inspired from the base-
line feature set suggested by the WMT2012 QE shared task
organisers (Callison-Burch et al., 2012). The full baseline
set initially contains 17 features, 2 of them being dependent
on the MT system. As Moses was the MT system used by

\[^4\text{https://svn.code.sf.net/p/apertium/svn/trunk/apertium-hbs-slv/}\]
\[^5\text{https://www.ciklopea.com}\]
the shared task organisers and the 2 system dependent features are extracted from GIZA word-alignment tables, we decide to exclude these features from our feature set and keep a baseline set as independent as possible from the MT system used.

The 15 baseline features are described below:
- 6 Surface Features source and target segment lengths, number of punctuation marks, average source word length and average target word occurrence.
- 2 Language Model Features 3-gram log-probabilities of source and target segments according to Kneser-Ney-discounted LMs built with the SRILM toolkit (Stolcke et al., 2011) using the slWaC and hrWaC monolingual corpora (Ljubešić and Erjavec, 2011) for source and target LMs respectively.
- 7 n-gram Frequency Features The number of source segment unigrams seen in a reference corpus (slWaC) plus 6 features based on the most and least frequent source n-gram (n ∈ {1; 3}) quartiles. The reference corpus is the corpus used to extract the LM features.

In order to improve the QE performance and to measure its impact on synthetic-data-based SMT, we extend the baseline feature set to 189 features including the baseline ones. This extended set contains:
- 36 Surface Features uppercased-lowercased letters ratio, untokenised items, special characters, source and target features ratio.
- 90 Language Model Features source and target 1 to 5-gram perplexity and log-probability according to LMs and backward-LMs (based on Raybaud et al. (2011)), as well as source and target features ratio.
- 63 n-gram Frequency Features source and target unigrams seen in a reference corpus (slWaC and hrWaC respectively), plus 1 to 5-gram frequencies in each of the frequency quartiles, as well as source and target features ratio.

3.3. Quality Estimation Evaluation
To evaluate the regression model, we infer TER scores at the sentence-level for each pair of the test set. The evaluation metrics are MAE, RMSE and Pearson’s r, but only the correlation coefficient is presented in this paper (Table 2). While MAE and RMSE are error measures (the lower the better) and thus indicate how far on average the predicted scores are from the reference ones, Pearson’s r is a correlation measure (the higher the better) and allows us to see whether the prediction follows a similar distribution to that of the reference. This latter score is the most interesting for us and we decide to select the best QE models based on this measure.

We build a regression model using each of the training corpora and evaluate them with the different test sets. This evaluation method aims to indicate which training corpus is performing the best on its corresponding test set, but also which corpus leads to a more generalised QE model. The

Pearson’s r results show that each of the training set performs best on its corresponding test set, while the slhr corpus leads to the best r score overall on its corresponding test set with QE model trained on the extended feature set. The data concatenation (system noted all) yields to a higher correlation score on the mixed test set and thus indicates a better generalisation over the training data. This motivates our choice to select this QE model for filtering translated monolingual data and generating a synthetic parallel corpus.

When comparing the baseline and the extended feature sets, we observe fluctuating improvements according to Pearson’s correlation coefficient regarding the training and testing corpora. Five data configurations lead to negative correlations when using the baseline features while it is not the case with the extended set. Using the EUb corpus for training and testing the QE model, extending the feature set does not lead to significant improvement (with p ≤ 0.01 using the bootstrap resampling method). For the other corpora, the extended feature set improves over the baseline set when the train and test sets are taken from the same corpus. Figure 1 shows the distributions of TER scores for the reference, the baseline and the extended QE setups, with the concatenated training and testing datasets (noted all). Better predictions are done by the extended QE model when the TER reference scores are low, while the baseline QE model tends to predict scores around the reference average. We decide to keep two QE models for the rest of our experiments, one using the baseline feature set and one using the extended set, both trained on the concatenated corpora.

4. Synthetic-data-based SMT
For the remaining experiments presented in this paper, the QE models are used individually to estimate TER scores at the sentence-level in order to filter translations provided by an RBMT system. The translations are then ranked according to their TER scores and subsets of this corpus are extracted to train SMT systems. These translation systems are finally evaluated with four of the most popular automatic metrics according to two test sets. Subsection 4.1.
and 20% of the training data. For the test set the extended QE models lead to the highest scores for 10

sion

The B

LEU

Denkowski, 2009).

13

test sets scored with B

The SMT systems are then evaluated based on the trans-

experiments three times and average the obtained results.

line, we randomly select subsets of the translated corpus

in the parallel corpus. To compare our approach to a base-

10

est TER score and we extract four subsets of this corpus,

Translations are then ranked from the lowest to the high-

ter side of the Europarl English–Slovene parallel corpus

into Croatian using the Apertium RBMT system. The re-

sulting English–Croatian parallel corpus is used to train a

phrase-based SMT system with the Moses tool-kit. We do

not run any tuning algorithm, and thus do not need a develop-

ment set, on the different SMT systems built in order to

strictly evaluate the effect of QE-based synthetic data gen-

eration. To evaluate the SMT systems, we use two different

test sets: Newstest2013 (a subset from WMT’13 test set

manually translated into Croatian) and SETimes. Details

about the training and testing datasets are presented in Ta-

ble 3.

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>621k</td>
</tr>
<tr>
<td>Slovene</td>
<td>621k</td>
</tr>
<tr>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>Newstest2013 (source)</td>
<td>1k</td>
</tr>
<tr>
<td>SETimes (source)</td>
<td>2k</td>
</tr>
</tbody>
</table>

Table 3: Number of sentences and words in the training and testing data used for the SMT system.

4.2. SMT Systems Evaluation

Based on the QE model presented in Section 3, we infer

TER scores at the sentence-level for each translated sen-

tence from the parallel training data presented in Table 3.

Translations are then ranked from the lowest to the high-

est TER score and we extract four subsets of this corpus,

keeping 10, 20, 40 and 80% of the overall amount of words

in the parallel corpus. To compare our approach to a base-

line, we randomly select subsets of the translated corpus

with similar amount of words. We repeat the random-based

experiments three times and average the obtained results.

The SMT systems are then evaluated based on the trans-

lated test sets scored with BLEU (Papineni et al., 2002)

version 13a, TER (using TERCeM) and METEOR (Lavie and

Denkowski, 2009).

The BLEU scores are presented in Figure 2, and show that

the extended QE models lead to the highest scores for 10

and 20% of the training data. For the test set SETimes, the

extended QE model also leads to the highest score for 40% of

the training data, while the baseline QE model is better

for this subset size on the Newstest2013 test set. This par-

ticular result can be explained by the fact that only a few

subset sizes are evaluated and the maximum BLEU score

obtained by the extended QE model may be higher than

the one obtained by the baseline QE model. Overall, the

two QE setups show better results compared to the random

setup for smaller training data. These results are explained

by the ability to select the best translations provided by the

RBMT first when using a QE-based approach compared to a

random selection of the translations.

Figure 2: BLEU scores obtained by the random and the two

QE setups on the two test sets, depending on training data

subset sizes.

The TER scores are presented in Figure 3 and show consis-

tent results based on the BLEU scores described previ-

ously. With 10 and 20% of the training data, the extended

QE model leads to lower TER scores compared to the

baseline QE model and the random approach. For New-

test2013, the lowest TER score is obtained by the ex-

tended QE model with 40% of the training data, while 80%

of the training data is necessary to obtain the lowest TER

score on SETimes with the QE models. For this latter test

set, the TER results are similar to the BLEU ones where

80% of the training data appears to lead to the best score,

once again explainable by the limited number of evaluated

subset sizes.

The METEOR scores are presented in Figure 4. For the

test set Newstest2013, the best METEOR score is obtained

by the extended QE model using 20% of the training data.

Increasing the training data subset size does not lead to an

Table 2: Pearson’s r obtained on the three corpora and the data concatenation using the baseline and the extended feature sets. Results in bold indicates the highest correlation between the prediction and the reference amongst training corpora for a given test set.

![Table 2](http://nlp.ffzg.hr/resources/corpora/)

setimes-hr
improvement of this result which indicates that no useful parallel data is found over 20% of the training data size. For the test set SETimes, the best METEOR score is obtained by the extended QE model using 10% of the training data. With 20 and 40% of the training data, the extended QE model still leads to the highest METEOR score compared to the baseline QE model and the random approach, while the baseline QE model is better than the two other systems using 80% of the training data which is similar to the results obtained on the Newstest2013 test set.

As shown by the evaluation done with three automatic metrics, the QE-based approach leads to better results with smaller amount of training data compared to the random selection of synthetic parallel instances. In order to validate these results, we perform statistical significance tests on BLEU between the random and the QE-based systems, using the paired bootstrap resampling method suggested by (Koehn, 2004). We use the toolkit provided by CMT9 which is based on the script mteval-v13a released by NIST10. We compare the extended QE-based approach with the three random systems individually (which were averaged previously to compute automatic metrics) considering two significance levels (p-values): 0.05 and 0.01. The results are presented in Table 4 and confirm our statement that the QE-based approach leads to better translations according to BLEU, compared to selecting random training instances, when the size of the training subset is below 40% of the synthetic training corpus.

As the amount of the synthetic training data increase, the performances of the random and QE-based systems become non-significantly different and the QE-based systems never outperform significantly the system trained using the full synthetic parallel corpus. It appears that the QE-based systems do no benefit from the remaining 80% of the QE-ranked parallel corpus. In order to verify that BLEU really reflects the translation quality, a native Croatian evaluator reviewed and assigned two scores at the sentence-level for the Newstest2013 corpus. The evaluator has access to the English source sentence and its translation performed by three SMT systems: the random and extended QE-based systems trained on 40% of the parallel data, as well as the systems trained on the full synthetic corpus. Each translation is evaluated on a 1 to 10 scale according to the fluency and adequacy criteria.

The results given by the human evaluator confirm what is observed using BLEU and only a few instances of Newstest2013 are better translated using the QE-based approach compared to the full system. Four examples of the QE-based system outperforming the two other ones are presented in Table 5 along with their fluency and adequacy scores.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Random1</th>
<th>Random2</th>
<th>Random3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newstest2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>p ≤ 0.01</td>
<td>p ≤ 0.01</td>
<td>p ≤ 0.01</td>
</tr>
<tr>
<td>20%</td>
<td>p ≤ 0.05</td>
<td>p ≤ 0.01</td>
<td>p ≤ 0.01</td>
</tr>
<tr>
<td>40%</td>
<td></td>
<td>p ≤ 0.05</td>
<td></td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td>p ≤ 0.01</td>
<td></td>
</tr>
<tr>
<td>SETimes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>p ≤ 0.01</td>
<td>p ≤ 0.01</td>
<td>p ≤ 0.01</td>
</tr>
<tr>
<td>20%</td>
<td>p ≤ 0.01</td>
<td>p ≤ 0.01</td>
<td>p ≤ 0.01</td>
</tr>
<tr>
<td>40%</td>
<td></td>
<td>p ≤ 0.01</td>
<td></td>
</tr>
<tr>
<td>80%</td>
<td></td>
<td>p ≤ 0.01</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Significance levels when comparing BLEU scores obtained by the extended QE-based system and the random systems. The p-values are calculated when the QE-based system reaches higher BLEU scores than the random systems.

9http://www.ark.cs.cmu.edu/MT/
As future work, we would like to investigate the use of a
quired to claim for the robustness of our approach.
results obtained in this study, and more experiments are re-
English–Slovene parallel corpus would also impact the re-
Improving the translation qual-
on automatic feature selection, could lead to some improve-
We assume that further improvements of the QE system,
This paper has presented a first step to the generation of
We show a significant improvement of the translation qual-
we depart from the synthetic corpus approach to
The case study presented deals with translation from
We have built a syn-
A QE system has been used to filter the resulting
QE setup are still unclear, for instance the performance of
more diverse feature set, containing linguistic information
such as part-of-speech and syntax, which were shown to
We assume that the full system is not significantly different
than the QE one because our approach quickly reaches a
plateau by using most of the good quality synthetic data in
the first 20%.

5. Conclusion
This paper has presented a first step to the generation of
parallel data for under-resourced languages using
QE. We departed from the synthetic corpus approach to
pivot-based MT and extended it by filtering the resulting
corpus with QE.

The research leading to these results has received fund-
ing from the European Association for Machine Trans-
lation through its 2011 sponsorship of activities program
and from the European Union Seventh Framework Pro-
gramme FP7/2007-2013 under grant agreement PIAP-GA-
2012-324414 (Abu-MaTran).

<table>
<thead>
<tr>
<th>Source</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>one thing is certain : these new provisions will have a negative impact on voter turn-out.</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>jedno je izvjesno : tim novim odredbama če imati okrnite udeleženost na izborima.</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>jedno je izvjesno : tim novim odredbama če imati negativan utjecaj na glasačko turn-out.</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>jedno je izvjesno : tim novim odredbama če imati okrnite udeleženost na izborima.</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>cigarettes are linked to 85 % of lung cancer cases.</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cigarette su povezane sa 85 % pljucnega rakavih slučaja.</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>cigarette su povezane sa 85 % pljucnega slučaja raka.</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>cigarettes navedenjena do 85 % pljucnega rakavih nepobolšanenkog.</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>however, in this vital area, much remains to be done.</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ali , u tom vitalnem cromane jo učiniti .</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>ali , u tom ključnom području , što jo treba učiniti .</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>ali , u tom ključnom području , dosta postoriti .</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>i am a hero of the last century , when culture meant something .</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ja sam junak iz posljednjih stoljeća , kad je kultura u mislima .</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>ja sam junak iz prošloga stoljeća , kad je kulturu značio nešto .</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>ja sam junak iz prošloga stoljeća , kad je kultura u mislima .</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>cigarettes are linked to 85 % of lung cancer cases .</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cigaretami navezenja do 85 % pljucnega rakavih nepobolšanenkog .</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>jedno je izvjesno : tim novim odredbama imati okrnite udeleženost na izborima .</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 5: Examples of source sentences and their translations obtained with the systems trained on the full synthetic corpus (noted Full), 20% of the synthetic data extracted randomly (noted Random) and with the extended QE approach (noted QE).

6. References
Banerjee, P., Rubino, R., Roturier, J., and van Genabith, J. (2013). Quality estimation-guided data selection for do-
tion with Pivot Languages. In *Proc. of the International Workshop on Spoken Language Translation*, pages 143–149, Hawaii, USA.
shop on statistical machine translation. In *Proceedings...*


