ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D4.3
Improved baseline SMT systems adjusted for narrow domain

Version No. 1.0
29/06/2012

Document Information

Deliverable number:	D4.3
Deliverable title:	Improved baseline SMT systems adjusted for narrow domain
Due date of deliverable:	31/03/2012 postponed to 30/06/2012
Actual submission date of deliverable:	29/06/2012
Main Author(s):	Sabine Hunsicker, Yu Chen
Participants:	Sabine Hunsicker, Yu Chen, Tilde
Internal reviewer:	Gregor Thurmair
Workpackage:	WP4
Workpackage title:	Comparable corpora in MT systems
Workpackage leader:	DFKI
Dissemination Level:	**PU**
Version:	V1.0
Keywords:	statistical machine translation, language resources, under-resourced languages , narrow domains

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
V0.1	01/06/2012	Draft	DFKI	Initial version prepared	Submitted for review
V0.3	26/06/2012	Draft	LT	Internal reviw	Submitted for final changes
V1.0	29/06/2012	Final	DFKI	Final changes	Submitted to PO

EXECUTIVE SUMMARY

Appropriate translations often depend on the domain of the texts we wish to translate. Often the available parallel corpora have a different domain than the one we would like to work on. In this deliverable we present our experiments with comparable corpora for narrow domains. Again the ACCURAT toolkit is used to gather comparable corpora and extract parallel data from them. Additionally we also make use of terminology lists which have been collected. Using this data, we extend the baseline SMT systems in order to improve the translation quality for the Renewable Energy domain.

Table of Contents

Abbreviations

Table 1. Abbreviations.

Abbreviation	Term/definition
MT	Machine Translation
SMT	Statistical Machine Translation
TM	Translation Model
LM	Language Model
BLEU	Bilingual Evaluation Understudy
AC	Acquis Communautaire
EU	European Union
TU	Translation Unit
WMT	Workshop for Statistical Machine Translation
MERT	Minimal Error Rate Training

Introduction

When translating in a narrow domain, it is particularly important to have in-domain parallel data, as, for example, lexical choices differ between different domains. Often we do not have enough in-domain data and only have access to general domain or out-of-domain data. In these cases we can try compensating by extracting data from in-domain comparable corpora or using additional in-domain terminology. In this deliverable we report on our experiments in the *renewable energy* domain. We tested different setups to make use of the additional data extracted from in-domain comparable corpora and also pre-processing the text to be translated by adding translations from a bilingual terminology list.

1. Methodology

In this section we describe how we adapt the baseline SMT systems to a narrow domain. We chose the *renewable energy* domain for our experiments, as partners were able to extract enough usable parallel data for this domain.

1.1. Using Parallel Data Extracted from Comparable Corpora

To make use of the extracted data, we again used the method of interpolating language models trained on the different corpora. This approach is described in detail in D4.2.

Similarly to adapting the general domain models, we train a language model on the target side part of the comparable corpus for the domain, which we then interpolate with the language model(s) trained on the baseline corpora. For the interpolation we use an in-domain development set. This ensures that the data from the in-domain language models gets the proper weighting.

We add the extracted parallel data to the baseline corpora and retrain the translation model.

1.2. Factored Translation Models

Phrase-based SMT only uses the surface forms to create the translation model. But especially when translating from or into a morphologically rich language such as Latvian, this creates problems. For instance, Latvian has two genders, two numbers and seven different cases for nouns, and as such the exact wordform might be unknown to our model, although we have seen the wordform before. Factored models try and fix this problem by using additional information about the surface forms, such as lemma or part of speech. The decoding step may then consist of several translation steps using different translation models.

1.3. Using Terminology Data

D2.3 reports the efforts of the consortium in extracting terminology and named entity data from the comparable corpora. We decided to also make use of this information.

One approach was to pre-process the test set with information from bilingual terminology lists. The source text is annotated with XML tags containing the matched terms from the list and a probability, which we take from the terminology lists.

2. Data

The baseline systems are identical with those listed in D4.2. In this section we will discuss the comparable narrow-domain corpora and the extracted pseudo-parallel data as well as the terminology data.

2.1. Comparable Corpora

Table 2 shows the data collected by ILSP for the renewable energy domain.

Table 2. Statistics of comparable corpora in renewable energy domain.

Language Pair	Size (lines)
Croatian	19,742
Lithuanian	62,902
Latvian	23,893
Romanian	39,671
English	607,816

After running the ACCURAT toolkit, we received the extracted corpora reported in
Table 3.

Table 3. Statistics about extracted corpora.

Language Pair	Size (lines)
Croatian-English	8,237
Lithuanian-English	16,743
Latvian-English	22,992
Romanian-English	26,939

As the amount of extracted data is rather small, we again use the full comparable corpora to train the language model.

2.2. Terminology Data

The comparable corpora were also used to extract term lists including named entities. For this, each monolingual corpus was tagged using terminology tools. The terms in the monolingual corpora were then later mapped to each other, resulting in a bilingual term list (see D2.3 for details on this). Table 4 shows a few examples.

Table 4. Terminology examples.

Latvian	English	Probability
Pakistāna	Pakistan	0.6659115852190215

Portugāle	Portugal	0.6659115852190215
Portugālē	Portugal	0.6659115852190215

This terminology data can be used to add additional translation options to the input text, which can then be used during decoding, as described in Section 3.3.

In total, our bilingual term lists contain 1836 terms for English—Latvian.

Additionally to the bilingual lists, we also have monolingual lists concerning many more term entries: 245972 English terms and 77623 Latvian terms. These were used for the factored translation models as described in Section 1.2.

2.3. Factored Data

To create the additional linguistic information for the factored translation models, we use a variety of tools. To get lemma and part of speech information for English, the TreeTagger was used. Terminology information was provided by the data given in Section 2.2. For Latvian, we used Tilde's Tagger Web Service based on a maximum entropy classifier[1]. The training corpora were tokenised, lemmatised, and morpho-syntactically tagged.

2.4. Development & Test Data

Table 5 reports the amount of development and test data available for the *renewable energy* domain for the four language pairs investigated.

Table 5. Statistics about development and test sets.

Language Pair	Development Set	Test Set
English-Croatian	500	500
English-Lithuanian	413	500
English-Latvian	526	1000
English-Romanian	500	663

[1] Pinnis, M., & Goba, K. (2011). Maximum Entropy Model for Disambiguation of Rich Morphological Tags. In C. Mahlow & M. Piotrowski (Eds.), *Proceedings of the 2nd International Workshop on Systems and Frameworks for Computational Morphology* (pp. 14-22). Zurich, Switzerland: Springer Berlin Heidelberg.

3. Experiments

We applied the approaches discussed in Section 1 for the following four language pairs:

- English→Croatian
- English→Lithuanian
- English→Latvian
- English→Romanian

Table 6. Statistics about training corpora.

Language Pair	Parallel Corpora	Size (lines)	Monolingual Corpora	Size (lines)
English→Latvian	DGT, JRC, ILSP-RenEn	2,328,666	DGT, JRC, ILSP-RenEn	2,329,567
English→Lithuanian	DGT, JRC, ILSP-RenEn	2,356,648	DGT, JRC, ILSP-RenEn	2,402,807
English→Croatian	SETimes, ILSP-RenEn	166,187	SETimes, ILSP-RenEn	177,692
English→Romanian	SETimes, ILSP-RenEn	198,512	SETimes, ILSP-RenEn	211,244

In Table 6 we see the total amount of training data available for each language pair. We have two language pairs of each "under-resourced" category: English→Latvian and English→Lithuanian have a large resource with the DGT/JRC corpus based on Acquis Communautaire, but it is heavily domain-dependent, namely of the legislation/law domain. The SETimes corpus, which forms the basis for English→Croatian and English→Romanian, is in the general domain (news text), but it is fairly small. In the following we describe the experiments: Section 3.1 reports the results when using the interpolated language model setup for all languages. The following two sections only deal with English→Latvian: in Section 3.2 we discuss the factored models and in Section 3.3 we use the terminology data via XML pre-processing.

3.1. Interpolated Language Models

Table 7. BLEU scores for interpolated LM experiments.

Language Pair	Baseline	Interpolated LM
English→Croatian	11.81	14.08 (+2.27)
English→Lithuanian	10.60	42.44 (+31.84)
English→Latvian	17.76	18.79 (+1.03)
English→Romanian	13.48	16.52 (+3.04)

Table 7 reports the BLEU scores for the interpolated language models, where we see improvements for all language pairs. We observe the largest gain in BLEU score for English→Lithuanian using DGT/JRC. Here the narrow domain corpus has the biggest impact. For English→Romanian, the improvement is still significant, although not quite as high.

To examine the difference in improvement, we calculated the amount of OOV (out of vocabulary) words for the baseline and enriched systems, i.e. we calculated how many tokens in the test set do not appear in the training corpus and thus cannot be translated properly. We do this for both the source text and the reference translation, as we also need to know if our model could calculate the appropriate translation. By adding additional data we hope to cover more of the tokens unknown to the baseline systems. The goal is to decrease the OOV count for both source text (we have translations for hitherto unknown words) and reference (the appropriate translations are also featured in the training corpus, so our model could create the proper entry in the phrase table).

Here we also have to take into account the different domains of the baseline corpora: legislation in the case of DGT/JRC, and news text for the SETimes corpus. Table 8 gives the exact numbers in percentages of unknown tokens/types. Tokens are all unknown tokens in total over the test set, whereas type only counts unique tokens.

Table 8. Statistics about Out Of Vocabulary words in test set (given as token/type).

Language Pair	Baseline		Interpolated LM	
	Source	Reference	Source	Reference
English→Croatian	11.5% / 3.4%	18.3% / 8.6%	9.4% / 2.5%	10.8% / 4.3%
English→Lithuanian	1.6% / 0.7%	2.7% / 0.9%	0.9% / 0.1%	1.3% / 0.4%
English→Latvian	4% / 0.8%	4.9% / 1.3%	2.7% / 0.3%	3.5% / 0.8%
English→Romanian	8.0% / 2.1%	23.0% / 12.9%	3.9% / 0.7%	14.7% / 5.2%

We see for English→Croatian the highest OOV rate for the source language. Whereas the rate drops significantly for the other three language pairs, the additional data only adds 1% of previously unknown source words. Although we obverse the similar rate of decrease concerning the target language for all language pairs, we cannot match up the added target tokens to the source tokens.

For English→Latvian, we observe a huge increase in translation quality, which we explain by the reduction of unknown source words. Only 0.1% of unknown types remain, less than 1% of the entire input text. At the same time, the amount of unknown target tokens is reduced in a similar fashion, so that we assume that the additional data, although it is smaller than e.g. the data for English→Latvian, adds exactly the phrases which we need to achieve a correct translation.

3.2. Factored Models

To make better use of the additional term information, we also used factored models. In this type of model, tokens are not only represented by their surface form, but additional (linguistic) information is added, such as lemma and part of speech. We only applied this approach for one language pair, English→Latvian.

We tested three setups, which we will call *term*, *lemma* and *ling* according to the type of factors they use. Table 9 gives the details of all setups.

Table 9. Information used in factored models.

Type	term	lemma	ling
Surface Form	Yes	Yes	Yes
Lemma	No	Yes	Yes
Part of Speech	No	Yes	Yes
Morphology	No	No	No
Term	Yes	No	Yes

In these models, the phrase table is trained on the factored input corpus and no further adjustments are made. The three setups we tried differ in the type and amount of information we used. In *term*, we translate the source surface form and source term information into the target surface form. The *lemma* model only uses the lemma and part of speech information, whereas the *ling* model maps the lemma, part of speech and term information of the source side to the target surface forms.
We tested our setups on English→Latvian.
Table 10 reports the results.

Table 10. BLEU scores for experiments with factored models.

System	BLEU
baseline	17.76
term	11.76 (-7.00)
lemma	11.91 (-6.85)
ling	11.72 (-7.04)

We see a lot of degradation for the simple models. This is also due to data sparseness as adding factors to the input text reduces the frequency of the individual token consisting of $factor_1|factor_2|...|factor_n$.

3.3. Terminology Data

Using the data described in Section 2.2, we can use an XML format to annotate our test set with the translation options we find in the bilingual terminology lists.
We will illustrate this using the examples from Table 4 for Pakistan and Portugal. For example, for the input phrase Pakistan, we create the following new token in the source text:

<xml translation=" Pakistāna" prob="0.6659115852190215">Pakistan</xml>

This token then replaces the original Pakistan in the source text, e.g.:
The name Pakistan literally means " Land of (the) Pure " in Urdu and Persian . becomes:
The name <xml translation=" Pakistāna" prob="0.6659115852190215">Pakistan</xml> literally means " Land of (the) Pure " in Urdu and Persian .

We can also use this format to annotate several possible translations, for example for Portugal:

```
<xml translation=" Portugāle|| Portugālē" prob="0.6659115852190215||0.6659115852190215">Portugal</xml>
```

The Moses SMT decoder allows us two options to use the XML: in the exclusive mode, only the given translations in the XML format are considered. When using the inclusive flag, these translation options will compete with the entries from the phrase table during decoding.

Table 11 lists the results for all experiments with this preprocessing for English→Latvian. The best results for each model have been marked in bold.

Table 11. BLEU scores for XML preprocessing experiments for English→Latvian.

XML Option	Baseline	Interpolated LM
No XML	**17.76**	**18.79** (+1.03)
XML inclusive	16.90	18.25 (+1.35)
XML exclusive	14.40	15.41 (+1.01)

The terminology lists don't help either model. Here we see that the XML processing turns out to be too restrictive and that the SMT model itself already retains important phrase information from the corpora.

One reason for the degradation could be that not all appropriate translation options were included in the terminology list, as it was rather small with not even 2,000 entries.

4. Conclusions

In this deliverable we reported the results of our experiments on the *renewable energy* domain. We see that interpolating language models achieve an increase in translation quality for the language pairs we investigated.

When using a bilingual terminology database, we need to take into account the quality and size of these terms. The terminology lists used in our experiments did not offer sufficient information to achieve a gain as illustrated by the difference between the XML exclusive and the no XML set up. By restricting ourselves to the terms contained in our bilingual list, we lose more than 3 BLEU points for each model.

As we can see from the evaluation of the factored models in

Table 10 using factors requires a lot of data to avoid data sparseness problems. A solution to avoid degradations could be to use generation models, for example, one could translate the source lemma to the target lemma and then generate the wordform by using the lemma and an additionally generated morphological analysis.

5. List of tables