DELIVERABLE

Project Acronym: iTranslate4

Grant Agreement number: 250405

Project Title: Internet Translators for all European Languages

---

4.2 Evaluation protocol

Revision: version 1

Authors:

Csaba Oravecz (Nyelvtudományi Intézet – Magyar Tudományos Akadémia, RIL)
Bálint Sass (Nyelvtudományi Intézet – Magyar Tudományos Akadémia, RIL)
László Tihanyi (Nyelvtudományi Intézet – Magyar Tudományos Akadémia, RIL)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Table of contents:

Deliverable 4.1.

Evaluation protocol

Introduction

Machine translation evaluation is a notoriously difficult problem. Several automatic evaluation metrics have been developed but their reliability and correlation with human judgements are open to much criticism. For this reason, large scale manual evaluation has started to gain preference over automatic evaluation metrics (see eg. Callison-Burch et al., 2008), and results of the latter approach are generally considered as capable only of an unreliable indication of the quality of an MT system. Furthermore, there is an observable tendency for these metrics to rank statistical systems consistently higher than their rule based counterparts regardless of the overall quality of the translations (Callison-Burch et al., 2010). In the iTranslate4 project, beside the automatic evaluation much emphasis has been put into developing a manual evaluation component as the core module of the evaluation framework which will include a substantial amount of user feedback from the translation portal to be collected in the second year of the project, on which a reliable ranking of translation alternatives could be based. At present, the results from the automatic evaluation framework and preliminary results of human evaluation as well as can only be reported. Consequently, no explicit ranking of translations are currently presented to the end user.

The automatic evaluation framework

Resources

Automatic evaluation methods are based on human reference corpora. In the ideal case 3 reference translations are taken into account, however, with such a wide variety of languages within the project to get hold of so many translations is completely unfeasible and so only 1 reference translation is used. It is assumed that variation in domains covered makes up for the lack of human translations. The required language resources have been collected from the EU news parallel corpus, with the following domains covered: agriculture, eu_explained, business, external_relations, culture, justice, economy, regions, employment, science, energy, transport, environment. The size of the resource is about 1900 paragpraphs per language of which only 25 were used for the first round of evaluation. Currently 21 languages are covered resulting in 56 language pairs (out of the 94 requested).

Description of the framework

The automatic evaluation framework is built around the IQMT toolkit (Giménez, 2007), which is a common workbench integrating a number of standard evaluation methods and metrics.

The input to the IQMT workbench was provided by a toolchain of wrapper scripts including a translation script using the standard APIs of the different translator engines to produce the translations and then filters to convert the text into the format required by IQMT. Evaluation results were calculated as a normalized average of 5 individual metrics: BLEU, NIST, GTM, METEOR and ROUGE. Various rankings have been prepared from the raw average, including a downweighted version of statistical translators with respect to rule based ones. However, due to the above mentioned reasons reliable rankings can only be acquired if extensive human evaluation results are available including user feedback data. From the three evaluation sources the final score can be calculated as a weighted average of the three components (automatic evaluation, Turkers' evaluation, user feedback):

$$score = (n_a * Q_a + n_h * Q_h + n_f * Q_f)/3$$

Specification of tools

Translator script

Purpose: The translator wrapper contacts all available translation services (TS) for each language pair to be evaluated and sends the input text to these services.

Usage: itrans_translate_xml.pl <input xml file> <output xml file> <target language> <TS>

Interfaces:
Input: A text file in mteval XML format in the source language.
Output: A text file in mteval XML format in the target language.

Input example:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "mteval-xml-v1.5a.dtd">

<mteval>
<srcset setid="itranslate4" srclang="en">

<doc docid="eunews_1" genre="nw">
<seg id="1">EU veterinary week focuses on identifying and tracking animals.</seg>
```

Output example:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "mteval-xml-v1.5a.dtd">

<mteval>
<tstset setid="itranslate4" srclang="en" trglang="de" sysid="XXX">
```

<doc docid="eunews_1" genre="nw">
<seg id="1" status="ok">ein eu den Woche Tierarzt nieten identifiziert und Biester nachsetzt.</seg>

Resources: The parallel language resources collected (xx language, xx language pairs).

IQMT driver script

Purpose: This script generates the input and configuration files needed by the IQMT workbench then runs the necessary IQMT evaluation steps and generates the required metrics.

Usage:

itrans_eval.sh [-h|t] [-p dir] [-a dir] <source> <target> <date>

Options:

-p path root directory of your itranslate repository. May be set
by the 'ITRANSLATE_HOME' environmental var.

-a path main directory for itranslate data, in which the
'ref, src, tst, results' directories are located.
May be set by the 'ITRANSLATE_DATA' environmental var.

-h this usage info

-d debug mode

-t test mode

Invocation:

Source and target languages must be specified by codes.

Legitimate language codes are the following:

af sq ar hy az eu be br bg ca zh hr cs da nl en eo et tl fi fr gl ka de el ht he hi hu is id ga it ja kk ko la lv lt mk ms mt no nn oc fa pl pt ro ru sr sk sl es sw sv th tr uk ur vi cy yi

Date format: YYMMDD

Interfaces:

Input: 1. A text file in mteval XML format in the source language. 2. A text file in mteval XML format in the target language containing the reference translation. 3. A text file in mteval XML format in the target language containing the translation from the translation service.

Output: A text file in system internal format containing the metrics for the translation service.

Input example:

1. (same as above)

2.

<?xml version="1.0" encoding="UTF-8"?>

<!DOCTYPE mteval SYSTEM "mteval-xml-v1.5a.dtd">

<mteval>
<refset setid="itranslate4" srclang="en" trglang="de">

<doc docid="eunews_1" genre="nw">
<seg id="1">Auf der europäischen Veterinärwoche geht es in erster Linie um die Kennzeichnung und Rückverfolgbarkeit von Tieren.</seg>

3. (same as above)
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "mteval-xml-v1.5a.dtd">

<mteval>
<tstset setid="itranslate4" srclang="en" trglang="de" sysid="XXX">

<doc docid="eunews_1" genre="nw">
<seg id="1" status="ok">ein eu den Woche Tierarzt nieten identifiziert und Biester nachsetzt.</seg>

Output example:

ID: en-de XXX eun 110309
SYS BLEU-1 BLEU-2 BLEU-3 BLEU-4 BLEUi-2 BLEUi-3 BLEUi-4 GTM-1 GTM-2 GTM-3 MTR-exact MTR-stem MTR-wnstm MTR-wnsyn NIST-1 NIST-2 NIST-3 NIST-4 NIST-5 NISTi-2 NISTi-3 NISTi-4 NISTi-5 RG-1 RG-2 RG-3 RG-4 RG-L RG-S* RG-SU* RG-W-1.2
S5 0.3208 0.1216 0.0588 0.0329 0.0461 0.0137 0.0057 0.2011 0.0770 0.0587 0.2169 0.2285 0.2293 0.2310 2.2280 2.3480 2.3557 2.3562 2.3562 0.1199 0.0077 0.0005 0.0000 0.2655 0.0527 0.0112 0.0038 0.2048 0.0645 0.0759 0.1023

Resources: Language resources collected. Resources must be placed in an appropriate directory structure: within a main directory, source text files in 'src', reference translations in 'ref', system translations in 'tst' directories. The output are placed in the 'results' directory. Within each directory, there is a separate subdirectory for an evaluation run on a particular date. The following naming conventions must be complied with to ensure smooth running of the script:

* date subdirectory: YYMMDD (eg. 110309)
* source files: sourcelangcode-targetlangcode_3lettercorpusname_src_YYMMDD.xml
  (eg. en-de_eun_src_110309.xml)
* reference translations:
  sourcelangcode-targetlangcode_3lettercorpusname_ref_YYMMDD.xml
  (eg. en-de_eun_ref_110309.xml)
* system translations:
  sourcelangcode-targetlangcode_3lettersystemcode_3ltrcorpusname_tst_YYMMDD.xml

(eg. en-de_XXX_eun_tst_110309.xml)

Text converter

Purpose: This script is a simple filter used by the IQMT driver program to convert from mteval XML format to clean text (one sentence perl line) format.

Interfaces:
Input: mteval XML text file
Output: Clean text file.

Statistics script

Purpose: This program collects the output metrics files from the IQMT evaluation and prepares summary statistics.

Usage: eval_stat.pl [-h|d] [-m m1,m2,...] [eval file(s)]

m measure comma separated list of measures to include in statistics. Default:
BLEUi-4 GTM-3 MTR-wnsyn NISTi-4 RG-4

d print out debugging info

h print out usage info

Interfaces:
Input: Evaluation results from the IQMT driver script.
Output: Summary statistics file.

Input example: (same as output of IQMT driver script)
Output example:

no-en XXX s07 110309 0.28112 BLEUi-4::0.2932 GTM-3::0.1810 MTR-wnsyn::0.6267 NISTi-4::0.0519 RG-4::0.2528

bg-en XXX eun 110309 0.24852 BLEUi-4::0.1904 GTM-3::0.1980 MTR-wnsyn::0.6419 NISTi-4::0.0143 RG-4::0.1980

pt-en XXX eun 110309 0.23232 BLEUi-4::0.1828 GTM-3::0.1922 MTR-wnsyn::0.6100 NISTi-4::0.0103 RG-4::0.1663

pt-en XXX eun 110309 0.2284 BLEUi-4::0.1646 GTM-3::0.1887 MTR-wnsyn::0.6108 NISTi-4::0.0130 RG-4::0.1649

Evaluator script

Purpose: This program extracts relevant statistics from the evaluation summary file produced in the previous step and outputs a ranking of translation systems sorted by laguage pairs. It is also possible to demote statistical translators with a predefined scaling factor.

Usage: eval_normal.pl [-h|d|f|e] [-t t1,t2,...] [-w translator table] [-s scale] [statfile]

```
 e           use exclusion list to exclude translators from stat.
 t translator override default exclusion list with comma separated list
             of translators
 w file      use translator table file to scale down translators listed
             there. Default scaling factor: 0.6
 s number    specify scaling factor. This option must be used together with '-w'
 d           print out debugging info
 h           print out usage info
```

Interfaces:
Input: Summary statistics from the statistics script.
Output: Ranked list of translation systems together with scores.

Input example: (same as output above)
Output example:

```
bg    en    XXX    0.4450
bg    en    XXX    0.3915
bg    en    XXX    0.2993
bg    en    XXX    0.1614
da    en    XXX    0.4562
da    en    XXX    0.3645
da    en    XXX    0.3539
da    en    XXX    0.2444
```

Main wrapper

Purpose: This simple shell wrapper takes only a date argument and runs all above evaluation steps for all language pairs for which there are available system translations for the specified date.

Human evaluation using Mechanical Turk

Using Turkers to evaluate translation quality is a fast and inexpensive approach, proven to be very useful and reliable (Callison-Burch, 2009). In the experiments, 30 medium-length sentences were collected representing a range of different topics for every source language. These sentences were then translated into the target language using the available automatic translators. Only those language pairs were considered for which there is a *direct* translation and there is *more than one provider* / translation engine available. This altogether amounts to 94 language pairs – 38 source languages.

The task of the evaluators was to rank the translations from 1 (best) to 5 (worst). The 30 sentences were divided up in to 6 groups (HITs in MTurk terminology), each containing 5 sentences and

allowing the evaluator to work only with 5 source sentences at a time. As an illustration, in the Swedish-English language pair (4 translators) the instructions for the evaluators were the following:

Rank Machine Translation Outputs
Instructions:

- You are shown 5 Swedish sentences, each followed by 4 English candidate translations.
- Your task is to rank the translations from best to worst (ties are allowed). A translation is considered better if it reflects the meaning of the original sentence better.
- Fluency in English is required. You must have the appropriate qualification to work on this HIT.
- Please evaluate all translations.

The order of the translated sentences were shuffled according to a simple, deterministic shuffling algorithm, which ensures that each translator appears in the first, second etc. position at an equal number of times.

To ensure the high quality of the evaluation, namely that only evaluators with appropriate language knowledge could work on the tasks, evaluators had to take a quick web-based test. In this test they answered 4 multiple choice questions and they were allowed to work in our evaluation campaign only if their solution was free from any mistakes. They were presented with four different translations (with errors in morphology, syntax or vocabulary) for every sentence and they had to select the best translation. The instructions for this qualification task were the following:

Test your knowledge of Swedish and English. Choose the best translation!
Be careful, you will get the qualification only if you do not make any mistakes.

During the campaign, every sentence was evaluated with 3 different evaluators. Evaluators received 15 to 30 US dollar cents for a HIT as payment (more difficult language pairs - for example where English was neither source nor target – were rewarded by higher payment). This way we obtained (30x3=) 90 scores for every automatic translator in a language pair. The final score was calculated in two ways, first, as a simple arithmetic mean and second as in the method used in the EuroMatrix project (Callison-Burch et al., 2009b), where a translator gained a point when it was evaluated better than (or equal with) another translator, and the translator with the most points won. In most cases these two metrics gave the same ordering.
The human evaluation campaign is ongoing, results covering all the planned (53) language pairs expected in the beginning of the second reporting period.

References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 70–106, Columbus, Ohio, June. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 286–295, Singapore, ACL.

Chris Callison-Burch, Philipp Koehn, Christof Monz and Josh Schroeder. 2009a. Findings of the 2009 Workshop on Statistical Machine Translation. In Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 1–28, Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder. 2009b. Evaluation Campaign. EuroMatrix Deliverable 1.2b.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki and Omar Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 17-53, Uppsala, Sweden, ACL

Jésus Giménez. 2007. IQMT. A Framework for Automatic Machine Translation Evaluation based on Human Likeness, Technical Manual, TALP Research Center

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
V1	25.02.2011	Csaba Oravecz	RIL	A technical document describing the evaluation method used at itranslate4.eu to assess the quality of translation services

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.