ACCEPT

SEVENTH FRAMEWORK PROGRAMME

THEME ICT-2011.4.2(a)

Language Technologies

ACCEPT

Automated Community Content Editing PorTal

www.accept.unige.ch

Starting date of the project: 1 January 2012

Overall duration of the project: 36 months

Baseline machine translation systems

Workpackage n°4	Name: Improving SMT
Deliverable n°4.1	Name: Baseline machine transla tion systems
Due date: 31 March 2012	Submission date: 30 March 2012

Dissemination level: PU

Organisation name of lead contractor for this deliverable: University of Edinburgh

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement n°288769.*

SEVENTH FRAMEWORK PROGRAMME

Overview

This deliverable consists of trained statistical machine translation systems, so the role of this accompanying document is just to give brief details on how to access the systems and notes on how the systems were trained.

Six baseline systems have been deployed: four using data from Symantec (English to French, German and Japanese, and French to English), and two using data from Translators Without Borders (French to and from English).

Accessing the Systems

The baseline systems can be accessed via a web interface, or programmatically. The web interface is at http://accept.statmt.org/demo, using the username 'accept' and password 'motelone'. The programmatic interface uses a similar API to google translate, and is best illustrated by the python program at the end of this document.

System Details

All the systems are phrase-based Moses systems, trained with the standard Moses pipeline. The translation and lexicalised reordering models were trained with a concatenation of all the available parallel data, whilst for the language model a separate model was trained on each corpus, with all models interpolated together minimising perplexity on the tuning set. The Moses tokenisation and casing tools were used, except for Japanese were the Kytea[1] segmenter was used in training.

The parallel data for the Symantec models consisted of translation memory data supplied by Symantec (containing product manuals, marketing content, knowledge base content and website content), supplemented with the WMT12[2] releases of europarl and news-commentary. For the language models, the target sides of all the parallel data were used, together with monolingual data from the Symantec forums. The monolingual data was not included in the English-German system as it was found not to improve the Bleu score. The tuning and test data for the Symantec systems (500 parallel sentences each) consisted of forum data which had been translated with google translate, then post-edited by a translator.

The TWB systems were trained on parallel data extracted from documents translated by TWB over the last 3 years. These were supplied as Word and Excel documents, so the text was first extracted by Acrolinx, before being sentence-aligned with hunalign[3]. The training data was again supplemented by europarl and news-commentary, and the language models were built from the target data. The tuning and test data (1000 parallel sentences each for each direction) was drawn randomly from the TWB training data, and the alignments were checked by Geneva.

The following table shows the Bleu scores of each of the systems on the current test set (note that these scores are not comparable across test sets):

1 http://www.phontron.com/kytea/

2 http://www.statmt.org/wmt12/

3 http://mokk.bme.hu/resources/hunalign/

System	Pair	Bleu
Symantec	English->French	36.14
	English->German	19.73
	English->Japanese	19.81
	French->English	42.41
TWB	English->French	24.40
	French->English	32.73

Appendix: Programmatic Access to Servers

The following python code requests an English-French translation from the Symantec baseline system. To use the TWB system, just replace the 'sb' in the url with 'tb'.

```
#!/usr/bin/env python
# coding=utf8
#
# Translates using the google api style interface
#
import json
import urllib

def main():
    urls = ["http://accept:motelone@accept.statmt.org/demo/translate.php"]
    for url in urls:
         print url
        source = "en"
        target = "fr"
        input_text = "I clearly stated in my earlier post this is what the tech guy did - and I reported his
exact steps ."
        params = urllib.urlencode({'v' : '1.0', 'ie' : 'UTF8', ¥
            'langpair' : '%s|%s' % (source,target),¥
            'system' : 'sb', 'q' : input_text})
        f = urllib.urlopen(url,params)
        line = f.readline()
        response = json.loads(line)
        if not response['responseData']:
            print "Error: ", response['responseDetails']
        else:
            print response['responseData']['translatedText']
if __name__ == "__main__":
    main()
```

Appendix: Data Sets for Each System

The data comes from three sources: project data contributed by Symantec and TWB, and data released for the WMT12 evaluation campaign. The first two are available from the Accept internal website[4], and the other from the WMT12 website[5]

English-French (Symantec)

Parallel training: symc_bip_06_en_fr.en.zip, symc_bip_07_en_fr.fr.zip, news-comentary-v7, europarl-v7

Monolingual training: symc_bip_12_fr_forum.tgz

Tuning and Test: symc_bip_01_devtest.en, symc_bip_03_devtest.fr

English-German (Symantec)

Parallel training: symc_bip_04_en_de.en.zip, symc_bip_05_en_de.de.zip, news-commentary-v7, europarl-v7

Monolingual training:

Tuning and Test: symc_bip_01_devtest.en, symc_bip_02_devtest.de

English-Japanese (Symantec)

Parallel training: symc_bip_08_en_ja.en.tgz, symc_bip_09_en_ja.ja.tgz

Monolingual training: symc_bip_14_ja_forum.tgz

Tuning and Test: symc_bip_01_devtest.en, symc_bip_00_devtest.ja

French-English (Symantec)

Parallel training: symc_bip_06_en_fr.en.zip, symc_bip_07_en_fr.fr.zip, news-commentary-v7, europarl-v7

Monolingual training: symc_bip_11_en_forum.tgz

Tuning and test: GenevaPostEdited

English-French (TWB)

Parallel training: TWB Data (train.fr-en and train.en-fr from the Feb 2012 release), news-comentary-v7, europarl-v7

Monolingual training:

Tuning and test: Selected and removed from TWB data, aligned by Geneva

4 https://plone2.unige.ch/accept/info/links-to-data

5 http://www.statmt.org/wmt12/

French-English (TWB)

Parallel training: TWB Data (train.fr-en and train.en-fr from the Feb 2012 release), news-comentary-v7, europarl-v7

Monolingual training:

Tuning and test: Selected and removed from TWB data, aligned by Geneva