

MT Praktikum - Evaluation

23. Mai 2017

Um die Aufgaben auszuführen, können Sie Ihre Daten in folgendem Verzeichnis speichern:
`/project/smtstud/ss17/systems/USERNAME/`

Wir werden uns einige der Übersetzungen anschauen, die beim Workshop for Statistical Machine Translation 2010 Shared Translation Task eingereicht wurden.
(<http://www.statmt.org/wmt10/translation-task.html>)

Im Verzeichnis: `/project/smtstud/ss17/data/wmt10-xlats/` liegen Übersetzungen von vier Deutsch-Englisch-Übersetzungssystemen (in jeweils zwei Formaten). Sie sind nach BLEU an deutlich unterschiedlichen Plätzen der Rangliste.

Die passenden Quelldaten sind hier:

`/project/smtstud/ss17/data/wmt10-xlats/ref/wmt10-newssyscombttest2010-src.de.sgm`

Bitte melden Sie sich nach Abschluss jeder Aufgabe, sodass wir ausschließen können, dass Sie mit fehlerhaften Dateien weitermachen.

1. Manual Evaluation

Unter dem folgenden Link können Sie an der Manuellen Evaluation teilnehmen (verwenden Sie maneval als Passwort):

<http://www.eSurveysPro.com/Survey.aspx?id=59d42ecb-8492-43c6-abd8-49ebc5f83549>

(a) Sehen Sie sich die vier Übersetzungen auf der Webseite an.

- Können Sie die schlechteste Übersetzung identifizieren?
- Welche halten Sie für die beste Übersetzung?

Bei Evaluationen gibt es (unter anderen) zwei Hauptmethoden, automatische Übersetzungen manuell bewerten zu lassen: Punkte für Adequacy und Fluency vergeben oder Ranking von Systemen.

- (b) Vergeben Sie Punkte für Adequacy (wie adäquat ist der Inhalt wiedergegeben) und Fluency (wie flüssig liest sich die Übersetzung?)
- (c) Vergeben Sie einen Rang (1-4) für jedes System
 - Finden Sie diese Aufgabe schwierig? Wieso?
 - Denken Sie heute haben alle dieselben Punkte und Rankings vergeben?
 - Wie geht man bei einer Evaluation mit diesem Problem um?

2. BLEU score

Kopieren Sie sich die Übersetzungen im sgml-Format der vier Systeme in Ihr Verzeichnis.

```
cp /project/smtstud/ss17/data/wmt10-xlats/*.sgm .
```

Das von NIST herausgegebene offizielle BLEU evaluation script finden Sie hier:

```
/project/smtstud/ss17/bin/mteval-v11a-cmufix.b.pl
```

Rufen Sie das Script mit der Option -h auf und sehen Sie sich die Ausgabe an.

- (a) Verwenden Sie das Script, um die BLEU und NIST Scores für alle vier Systeme zu berechnen.

reference: /project/smtstud/ss17/data/wmt10-xlats/ref/wmt10-newssyscombttest2010-ref.en.sgm

source: /project/smtstud/ss17/data/wmt10-xlats/ref/wmt10-newssyscombttest2010-src.de.sgm

test: alle vier *.sgm files (aus: /project/smtstud/ss17/data/wmt10-xlats/)

Verwenden sie folgende Optionen für alle evaluation runs:

- i. Berechnen Sie BLEU und NIST Score
 - ii. Berechnen Sie den Score case insensitive
 - iii. Verwenden Sie die BLEU text normalization method
- Wie hoch ist der BLEU und NIST Score für die vier Systeme?
 - Was bedeuten Precision und Length Penalty des BLEU scores?
 - Sieht das automatische Ranking so aus, wie Sie es erwartet haben?
 - Wie wird der BLEU Score aus den n-gram matches berechnet?
 - Was ist der Unterschied zwischen dem Individual und dem Cumulative score?

System	NIST	BLEU	Prec.	LP
DFKI				
JHU				
KIT				
RWTH				

(b) Evaluieren Sie unser System (kit) mit der Option mit Unterscheidung von Groß- und Kleinschreibung.

- Wie unterscheidet sich der Score und warum?

(c) Evaluieren Sie unser System (kit) nur für BLEU auf Dokumentlevel:

```
/project/smtstud/ss17/bin/mteval-v11a-cmufix_b.pl  
-r /project/smtstud/ss17/data/wmt10-xlats/ref/wmt10-newssyscombttest2010-ref.en.sgm  
-s /project/smtstud/ss17/data/wmt10-xlats/ref/wmt10-newssyscombttest2010-src.de.sgm  
-t ./newssyscombttest2010.de-en.kit.sgm  
-b -m b -d 1 > & kit.doc-eval.log
```

- Wie kann es sein, dass ein Dokument einen Score von Null hat?

Finden Sie das Dokument mit dem höchsten und mit dem niedrigsten BLEU Score (ignorieren Sie das Dokument mit Score Null) und lesen Sie die Übersetzung.

- Lässt sich der unterschiedliche Score nachvollziehen?
- Spekulieren Sie über die Ursachen für die unterschiedliche Qualität der Übersetzungen.

3. TER Score

TER steht für Translation Error Rate. Die TER ist das Verhältnis von Fehlern in der Hypothese zur Anzahl von Wörtern in der Referenzübersetzung. Die Fehler setzen sich aus Insertions, Deletions, Substitutions und Shifts von Wörtern zusammen.

(a) Berechnen Sie die Translation Error Rate für den ersten Satz der Übersetzung unseres Systems:

- Wie viele Editieraktionen brauchen Sie, um aus der Hypothese die Referenz zu machen?

ref:	lack of snow	in	mountains	causes problems	for hoteliers
hyp:	lack of snow	on the	mountains	has bedeviled	the hotel owners

Rufen Sie das Programm zur Berechnung des TER Scores für diese Übersetzung auf:

```
/project/smtstud/ss17/bin/tercom.v6b.pl -h <hypothesis-file> -r <reference-file> -i sgm -N & (Laufzeit c.a. 5Min)
```

(b) Noch während das Programm läuft, können Sie sich die Datei **newssyscombttest2010.de-en.kit.sgm.sys.pra** einmal ansehen.

- Haben Sie die TER des ersten Satzes richtig berechnet?
- Ist es möglich dass die Fehlerrate auf über 100% steigt? Wie kann das passieren?

(c) Finden Sie Satz ID „idnes.cz/2009/12/11/76492-24“ in der Datei **newssyscombttest2010.de-en.kit.sgm.sys.pra**.

Der Quellsatz war: „Wir liegen etwa um fünf Prozent besser als im Vorjahr.“

- Wie hoch ist die TER des Satzes?
- Welche Fehlerrate hätten Sie (als Mensch) vergeben, wenn Sie sich nicht die Referenz sondern den Quellsatz zum Vergleich ansehen?
- Wie kann man diesem Problem teilweise entgegenwirken?

(d) In der Datei **newssyscombttest2010.de-en.kit.sgm.sys.sum.doc** finden Sie die TER scores für alle Dokumente und für das gesamte Testset.

- Was ist der TER Score für das kit System?

4. TER Score - Character based

Schauen Sie sich die folgenden Dateien an. Die Dateien enthalten die gleichen Sätze aber in einem anderen Format.

Referenz:

```
/project/smtstud/ss17/data/wmt16-ende/newstest2015-ende-ref.de.sgm
/project/smtstud/ss17/data/wmt16-ende/newstest2015-ende-ref.de
```

Hypothese:

```
/project/smtstud/ss17/data/wmt16-ende/newstest2015-ende.kit.de.sgm
/project/smtstud/ss17/data/wmt16-ende/newstest2015-ende.kit.de
```

Kopieren Sie sie in ihr Verzeichniss.

```
cp /project/smtstud/ss17/data/wmt16-ende/newstest2015-ende-ref.de.sgm .  
cp /project/smtstud/ss17/data/wmt16-ende/newstest2015-ende-ref.de .  
cp /project/smtstud/ss17/data/wmt16-ende/newstest2015-ende.kit.de.sgm .  
cp /project/smtstud/ss17/data/wmt16-ende/newstest2015-ende.kit.de .
```

- (a) Berechnen Sie die Übersetzungsqualität mittels der character-based TER mit dem folgendem Befehl:

```
python /project/smtstud/ss17/bin/CharacTER.py -r newstest2015-ende-ref.de  
-o newstest2015-ende.kit.de -v > wmt15_kit_charTER.Log
```

Schauen Sie sich die Ausgabedatei an.

- Welchen char-TER score hat das Dokument?

- (b) Berechnen Sie die Übersetzungsqualität mittels word-based TER mit dem folgenden Befehl:

```
perl /project/smtstud/ss17/bin/tercom.v6b.pl -r newstest2015-ende-ref.de.sgm  
-h newstest2015-ende.kit.de.sgm -i sgm -N
```

Schauen Sie sich folgende Datei an `newstest2015-ende.kit.de.sgm.sys.pra` und `newstest2015-ende.kit.de.sgm.sys.sum.doc`.

- Welchen wort-TER score hat das Dokument?

- (c) Vergleichen Sie character-based TER und word-based TER für jeden Satz

- Vergleichen Sie die TER für word-based and char-based für den ersten Satz. Welche ist höher? Wieso?

Ref: Die Premierminister Indiens und Japans trafen sich in Tokio.

Hyp: Indiens und Japans Ministerpräsidenten treffen sich in Tokio

- Vergleichen Sie die TER für word-based and char-based für Satz 380. Welche ist höher? Wieso?

Ref: Plötzlich war ich im Nationaltheater und ich konnte es kaum glauben.

Hyp: Plötzlich bin ich am National Theatre, und ich war einfach nicht ganz glauben.