Karlsruher Institut für Technologie

Karlsruhe Institute of Technology
Ngoc-Quan Pham
Institute for Anthropomatics and Robotics, Interactive Systems Labs
ngoc.pham@kit.edu

MT Praktikum - Evaluation Solution

26. Juni 2018

Um die Aufgaben auszuführen, können Sie Ihre Daten in folgendem Verzeichnis speichern: `/project/smtstud/ss18/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/ss18/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/ss18/data/wmt10-xlats/ref/wmt10-newssyscombtest2010-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

   Die Ergebnisse werden später bekannt gegeben.

   (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?

Antworten zu Aufgabe 1:

1.b) & c) Hier gibt es keine richtige oder falsche Antwort, manuelle Evaluation liegt im Ermessen des Bewertenden. Es ist schwierig, da Übersetzungsfehler schwer zu vergleichen sind und da ähnlich falsche Übersetzungen schwer zu ranken sind. Da oft sehr verschiedene Punkte oder Rankings vergeben werden, sammelt man für jeden Satz mehrere menschliche Bewertungen und misst die Übereinstimmungen.

2. **BLEU score**

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

```
cp /project/smtstud/ss18/data/wmt10-xlats/*.sgm .
```

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

`/project/smtstud/ss18/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/ss18/data/wmt10-xlats/ref/wmt10-newssyscombtest2010-ref.en.sgm
source: /project/smtstud/ss18/data/wmt10-xlats/ref/wmt10-newssyscombtest2010-src.de.sgm
test: alle vier *.sgm files (aus: `/project/smtstud/ss18/data/wmt10-xlats/`)

Verwenden sie folgende Optionen für alle evaluation runs:

i. Berechnen Sie BLEU und NIST Score
ii. Berechnen Sie den Score case insensitiv

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	5.9481	0.1767	0.1767	1.0
JHU	6.6869	0.2141	0.2274	0.9414
KIT	6.8816	0.2397	0.2397	1.0
RWTH	6.9411	0.2435	0.2435	1.0

(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/ss18/bin/mteval-v11a-cmufix_b.pl
-r /project/smtstud/ss18/data/wmt10-xlats/ref/wmt10-newssyscombtest2010-ref.en.sgm
-s /project/smtstud/ss18/data/wmt10-xlats/ref/wmt10-newssyscombtest2010-src.de.sgm
-t ./newssyscombtest2010.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.

Antworten zu Aufgabe 2:

2.a)

Die Precision ist der n-gram match count im Verhältnis zur Länge der Übersetzung. Unsinnig kurze Übersetzungen würden eine höhere Precision bekommen, daher werden

```
/project/smtstud/ss18/bin/mteval-v11a-cmufix_b.pl
-r /project/smtstud/ss18/data/wmt10-xlats/ref/wmt10-newssyscombtest2010-ref.en.sgm
-s /project/smtstud/ss18/data/wmt10-xlats/ref/wmt10-newssyscombtest2010-src.de.sgm
-t /project/smtstud/ss18/data/wmt10-xlats/newssyscombtest2010.de-en.kit.sgm
-m b
```

zu kurze Übersetzungen mit der Length Penalty bestraft. Sie sollte daher eigentlich Brevity Penalty heißen. Sie berechnet sich aus dem Verhältnis der Länge der Übersetzung zur Länge der Referenz, die der Übersetzung am nächsten ist. Ist die Übersetzung länger als die Referenz, ist die penalty 1 (= keine penalty).

Der BLEU score ist das geometrische Mittel der n-gram matches für die Längen 1 bis 4. Im Individual score ist der n-gram match score für jede n-gram Länge einzeln angegeben, bei Cumulative Score ist bereits das geometrische Mittel für alle kürzeren n-grams berechnet.

2.b)
NIST score = 6.6975 LP: 1.0000 BLEU score = 0.2288 Prec: 0.2288 LP: 1.0000 for system kit Der score ist niedriger, da jetzt Wörter nicht mehr matchen, bei denen nur die Klein- oder Großschreibung nicht stimmt.

2.c)
Ein Score von Null kommt zu Stande, wenn z.B. kein 4gram aus der Übersetzung in der Referenz vorkommt. Weil der BLEU das geometrische Mittel aus allen n-gram matches ist, wird der gesamte Score Null, wenn ein count Null ist.

Best: economist/2009/12/10/17240
Worst: elmundo/2009/12/11/10121

Im gut übersetzten Artikel gibt es viele kurze einfache Sätze. Im Artikel mit dem schlechten Score sind sehr viele unbekannte Wörter enthalten, die deutsch durchgereicht wurden.

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?

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

ref:	lack of snow	in	mountains	causes problems	for hoteliers
	0	1+1	0	2	2 + 1
hyp:	lack of snow	on the	mountains	has bedeviled	the hotel owners

/project/smtstud/ss18/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 **newssyscombtest2010.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 **newssyscombtest2010.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 newssyscombtest2010.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?

Antworten zu Aufgabe 3

3.a) 5 substitutions + 2 insertions / 9 words in the reference = 77.78%

3.b) Z.B. Satz ID idnes.cz/2009/12/11/76492-13 hat eine TER von 107.69%. Dies kann passieren, wenn die Hypothese länger ist als die Referenz und auch, wenn viele Shifts vorkommen. TER bevorzugt kürzere Hypothesen, was zu Problemen führen kann, wenn man ein SMT System auf TER hin optimiert.

3.c) Die automatische Übersetzung ist komplett richtig, bekommt aber eine katastrophale TER von 183.33, weil die Referenzübersetzung den Inhalt anders ausdrückt und zudem viel kürzer ist. Man würde hier als Mensch also 0 Fehler vergeben.

Man kann diesem Problem entgegenwirken, indem man z.B. den Text normalisiert (we're = we are, 5 = five, per cent = percent). Es kann auch sehr hilfreich sein,

wenn man mehrere Referenzübersetzungen für jeden Satz hat, sodass man verschiedene Weisen etwas auszudrücken zur Auswahl hat.

Sentence ID: idnes.cz/2009/12/11/76492-24
Source: Wir liegen etwa um fünf Prozent besser als im Vorjahr.“
Best Ref: we’re up about 5 percent .
Orig Hyp: we are about five per cent better than in the previous year .

REF: WE’RE UP about **** *** **** ****** **** ** *** 5 PERCENT .
HYP: WE ARE about FIVE PER CENT BETTER THAN IN THE PREVIOUS YEAR .
EVAL: S S I I I I I I I S S
SHFT:
TER Score: 183.33 (11.0/ 6.0)

3.d) TER 59.187

4. **TER Score - Character based**

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

Referenz:
/project/smtstud/ss18/data/wmt16-ende/newstest2015-ende-ref.de.sgm
/project/smtstud/ss18/data/wmt16-ende/newstest2015-ende-ref.de

Hypothese:
/project/smtstud/ss18/data/wmt16-ende/newstest2015-ende.kit.de.sgm
/project/smtstud/ss18/data/wmt16-ende/newstest2015-ende.kit.de

Kopieren Sie sie in ihr Verzeichniss.
cp /project/smtstud/ss18/data/wmt16-ende/newstest2015-ende-ref.de.sgm .
cp /project/smtstud/ss18/data/wmt16-ende/newstest2015-ende-ref.de .
cp /project/smtstud/ss18/data/wmt16-ende/newstest2015-ende.kit.de.sgm .
cp /project/smtstud/ss18/data/wmt16-ende/newstest2015-ende.kit.de .

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

Schauen Sie sich die Ausgabedatei an.

python /project/smtstud/ss18/bin/CharacTER.py -r newstest2015-ende-ref.de -o newstest2015-ende.kit.de -v > wmt15_kit_charTER.Log

- Welchen char-TER score hat das Dokument. Welcher Satz hat den besten und schlechtesten char-TER score? Wie hoch ist es?

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

perl /project/smtstud/ss18/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.

Antworten zu Aufgabe 4

4.a)
Dokument char-TER: 0.5674
Dokument wort-TER: 0.5849

4.c) First example:
char-based: 0.6889

word-based: 0.50
In word-based TER the sentence has a better performance. Second example:

char-based: 0.4205
word-based: 0.7500
In character-based TER the sentence is measured to be better matching to the reference.