Evaluation

Maschinelle Übersetzung

Samuel Läubli

Institut für Computerlinguistik
Universität Zürich

27. Februar 2018

Universität Zürich[UZH]

Themen

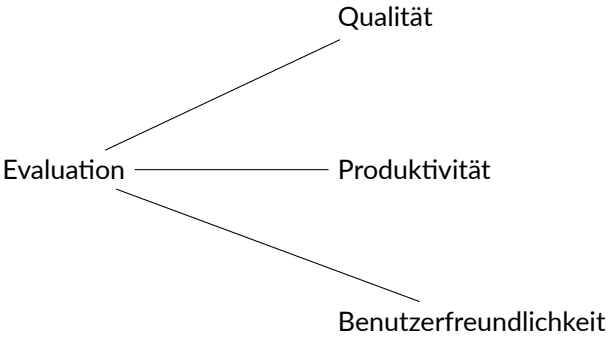
1. Einführung

2. Manuelle Evaluation

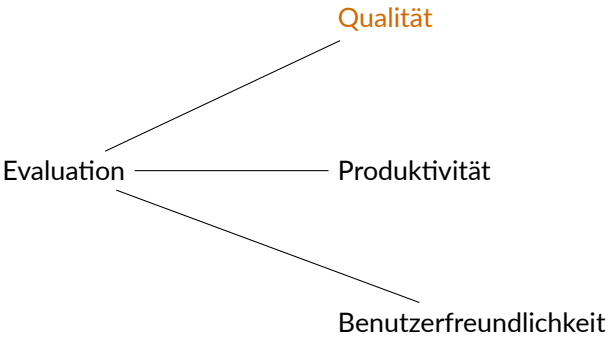
3. Automatische Evaluation

4. Zusammenfassung

Was bewerten wir?



Was bewerten wir?



Übersetzungsübung

The world is a stage, but the play is badly cast.

– Oscar Wilde

Qualitätsevaluation: Anforderungen

Eine Metrik zur Bewertung von Übersetzungsqualität soll folgende Anforderungen erfüllen:

Qualitätsevaluation: Anforderungen

Eine Metrik zur Bewertung von Übersetzungsqualität soll folgende Anforderungen erfüllen:

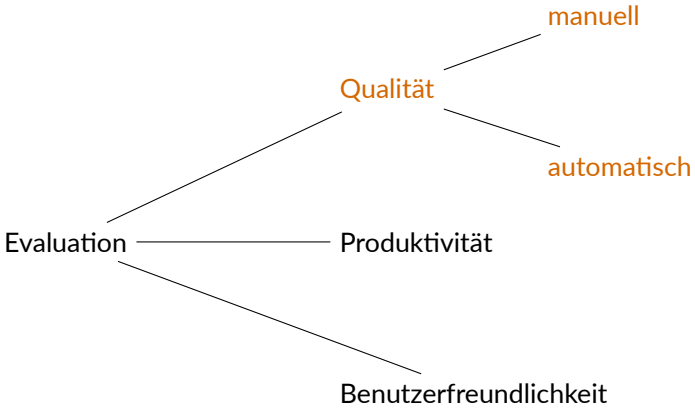
- **geringe Kosten**: Bewertung soll schnell und günstig sein
- **aussagekräftig**: Metrik soll einfach interpretierbar sein
- **konsistent**: Bewertungen sollen immer zum gleichen Ergebnis führen
- **korrekt**: Bewertung soll wahrheitsgemäss sein.

Qualitätsevaluation: Anforderungen

Eine Metrik zur Bewertung von Übersetzungsqualität soll folgende Anforderungen erfüllen:

- **geringe Kosten**: Bewertung soll schnell und günstig sein
- **aussagekräftig**: Metrik soll einfach interpretierbar sein
- **konsistent**: Bewertungen sollen immer zum gleichen Ergebnis führen
- **korrekt**: Bewertung soll wahrheitsgemäss sein. → Problem: Subjektivität.
  Es gibt keine (singuläre) «Wahrheit» (ground truth) in der Übersetzung.

Wie bewerten wir Qualität?



Vor- und Nachteile

Manuelle Evaluation

- \+ verlässlich(er)
- – teuer
- – langsam

Automatische Evaluation

- – unverlässlich(er)
- \+ billig
- \+ schnell

Themen

1. Einführung

2. Manuelle Evaluation

3. Automatische Evaluation

4. Zusammenfassung

Themen

1. Einführung

2. Manuelle Evaluation

3. Automatische Evaluation

4. Zusammenfassung

Beispiel

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

Beispiel

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

$\rightarrow$ Wie gut ist diese Übersetzung?

Beispiel

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

Auf einer Skala von 1–5,

- wie **adäquat** ist ist diese Übersetzung? (inhaltliche Korrektheit)
- wie **flüssig** ist diese Übersetzung? (Grammatikalität, Idiomatizität)

Beispiel

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

DeepL:
Die Welt ist eine Bühne, aber das Stück ist schlecht besetzt.

Beispiel

Original:
The world is a stage, but the play is badly cast.

Google Translate:
Die Welt ist eine Bühne, aber das Spiel ist schlecht besetzt.

DeepL:
Die Welt ist eine Bühne, aber das Stück ist schlecht besetzt.

Welche Übersetzung ist besser?

- Google Translate > DeepL
- Google Translate = DeepL
- Google Translate < DeepL

Absolute manuelle Evaluation

Maschinelle Übersetzungen können mit absoluten Werten beurteilt werden. Traditionell beurteilen wir **Adäquatheit** und **Flüssigkeit** auf einer 5-Punkte-Likert-Skala.

Absolute manuelle Evaluation

Maschinelle Übersetzungen können mit absoluten Werten beurteilt werden. Traditionell beurteilen wir **Adäquatheit** und **Flüssigkeit** auf einer 5-Punkte-Likert-Skala.

$\rightarrow$ Was bedeutet eine Flüssigkeit von 4?

Absolute manuelle Evaluation: Beispiel (WMT 2006)

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Quelle: Koehn und Monz, 2006

Absolute manuelle Evaluation

Adequacy:

5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency:

5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

Quelle: Koehn und Monz, 2006

Absolute manuelle Evaluation

Adequacy:

5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

Fluency:

5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

→ Was ist der Unterschied zwischen «much meaning» und «most meaning»?

Quelle: Koehn und Monz, 2006

Absolute manuelle Evaluation: Probleme

- Unklare Definitionen
- Unterschiedliche Personen vergeben unterschiedliche Durchschnittswerte
- Gleiche Person kann eigene Bewertung u.U. nicht reproduzieren
- Bewertung von Adäquatheit und Flüssigkeit korreliert stark; die Kriterien sind schwer auseinanderzuhalten

Relative manuelle Evaluation

Bewertungen sind i.d.R. konsistenter, wenn zwei oder mehr Systeme miteinander verglichen statt mit absoluten Werten versehen werden.

Relative manuelle Evaluation: Beispiel (WMT 2013)

For each ranking task, the judge is presented with a source segment, a reference translation, and the outputs of five systems (anonymized and randomly-ordered). The following simple instructions are provided:

You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed).

Quelle: Bojar et al., 2013

Relative manuelle Evaluation: Beispiel (WMT 2013)

Quelle: Bojar et al., 2013

Relative manuelle Evaluation: Pairwise Ranking

Aus relativen Evaluationen ergeben sich Bewertungen von Systempaaren A, B:

A besser als B	unentschieden	B besser als A
41	12	59

Relative manuelle Evaluation: Pairwise Ranking

Aus relativen Evaluationen ergeben sich Bewertungen von Systempaaren A, B:

A besser als B	unentschieden	B besser als A
41	12	59

$\rightarrow$ Ist System A besser als System B, oder ist der Unterschied zufällig?

Relative manuelle Evaluation: Pairwise Ranking – Signifikanz

Nullhypothese: Qualitätsunterschied zwischen System A und B resultiert aus zufälliger Variation.

Alternativhypothese: Qualitätsunterschied zwischen System A und B ist nicht zufällig.

Um die Nullhypothese zu verwerfen, erwarten wir

- weniger als 5% Wahrscheinlichkeit, dass Unterschied durch zufällige Variation bedingt ist $\rightarrow$ Unterschied mit statistischer Signifikanz von 95% ($p < 0.05$)

oder

- weniger als 1% Wahrscheinlichkeit, dass Unterschied durch zufällige Variation bedingt ist $\rightarrow$ Unterschied mit statistischer Signifikanz von 99% ($p < 0.01$)

Relative manuelle Evaluation: Pairwise Ranking – Signifikanz

Statistische Signifikanz kann mit dem Vorzeichentest (sign test) geprüft werden.
Beispiel in R:

```
> binom.test(59, 100, p=0.5, alternative="two.sided")

  Exact binomial test

data: 59 and 100
number of successes = 59, number of trials = 100, p-value = 0.08863
alternative hypothesis: true probability of success is not equal to 0.5
```

...

Relative manuelle Evaluation: Pairwise Ranking – Signifikanz

Aus relativen Evaluationen ergeben sich Bewertungen von Systempaaren A, B:

A besser als B	unentschieden	B besser als A
41	12	59

→ Ist System A besser als System B, oder ist der Unterschied zufällig?

Relative manuelle Evaluation: Pairwise Ranking – Signifikanz

Aus relativen Evaluationen ergeben sich Bewertungen von Systempaaren A, B:

A besser als B	unentschieden	B besser als A
41	12	59

→ Ist System A besser als System B, oder ist der Unterschied zufällig?
→ Qualitätsunterschied ist nicht statistisch signifikant, also zufällig.

Themen

1. Einführung

2. Manuelle Evaluation

3. Automatische Evaluation

4. Zusammenfassung

Daten

Wir teilen alle verfügbaren Übersetzungen in drei parallele Korpora auf: ein Trainings-, ein Validierungs- und ein Testkorpus. Es gelten folgende Grundsätze:

- Grösse Testkorpus: 1’000 bis 2’000 Sätze
- Sätze zufällig auswählen (!)
- Automatische Evaluation während Systementwicklung
- Nach Möglichkeit manuelle Evaluation vor Inbetriebnahme

Grundlage

Wie bewerten wir eine Übersetzung automatisch?

Eine Automatische Evaluationsmethode ist eine Funktion σ, welche die Ähnlichkeit zwischen einer maschinellen Übersetzung («Hypothese») h und Referenzübersetzung(en) r berechnet:

$$\text{score} = \sigma(h, r) \tag{1}$$

Die Ähnlichkeit wird liegt normalerweise zwischen 0.0 und 1.0 bzw. 0 und 100 %.

Voraussetzungen

- Ähnlichkeitsfunktion σ («Metrik»)
- 1..*n* Referenzübersetzungen für jeden zu bewertenden Satz

Precision, Recall, F-Measure

- **Precision** = $\frac{\text{korrekt}}{\text{hyp-länge}}$
  Wieviele Wörter der Hypothese sind in der Referenzübersetzung enthalten?
- **Recall** = $\frac{\text{korrekt}}{\text{ref-länge}}$
  Wieviele Wörter der Referenzübersetzung sind in der Hypothese enthalten?
- **F1-Measure** = $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
  Harmonisches Mittel von Precision und Recall.

Precision, Recall, F-Measure: Beispiel

Hypothese:
Israeli officials responsibility of airport safety

Referenz:
Israeli officials are responsible for airport security

$$\text{Precision} = \frac{\text{korrekt}}{\text{hyp-länge}} =$$

$$\text{Recall} = \frac{\text{korrekt}}{\text{ref-länge}} =$$

$$\text{F1-Measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} =$$

Precision, Recall, F-Measure: Beispiel

Hypothese:
Israeli officials responsibility of airport safety

Referenz:
Israeli officials are responsible for airport security

$$\text{Precision} = \frac{\text{korrekt}}{\text{hyp-länge}} = \frac{3}{6} = 0.5 = 50.0\,\%$$

$$\text{Recall} = \frac{\text{korrekt}}{\text{ref-länge}} =$$

$$\text{F1-Measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} =$$

Precision, Recall, F-Measure: Beispiel

Hypothese:
Israeli officials responsibility of airport safety

Referenz:
Israeli officials are responsible for airport security

$$\text{Precision} = \frac{\text{korrekt}}{\text{hyp-länge}} = \frac{3}{6} = 0.5 = 50.0\,\%$$

$$\text{Recall} = \frac{\text{korrekt}}{\text{ref-länge}} = \frac{3}{7} = 0.429 = 42.9\,\%$$

$$\text{F1-Measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} =$$

Precision, Recall, F-Measure: Beispiel

Hypothese:
Israeli officials responsibility of airport safety

Referenz:
Israeli officials are responsible for airport security

$$\text{Precision} = \frac{\text{korrekt}}{\text{hyp-länge}} = \frac{3}{6} = 0.5 = 50.0\,\%$$

$$\text{Recall} = \frac{\text{korrekt}}{\text{ref-länge}} = \frac{3}{7} = 0.429 = 42.9\,\%$$

$$\text{F1-Measure} = 2 \cdot \frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} = 2 \cdot \frac{0.5 \cdot 0.429}{0.5+0.429} = 2 \cdot \frac{0.214}{0.929} = 0.461 = 46.1\,\%$$

Precision, Recall, F-Measure: Problem

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

Precision =

Precision, Recall, F-Measure: Problem

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

Precision = $100.0\,\%$

Precision, Recall, F-Measure: Problem

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

Precision = $100.0\,\%$ $\rightarrow$ Wortstellung wird nicht berücksichtigt

Word Error Rate (WER)

Minimale Editierdistanz (Levenshtein-Distanz) von Hypothese zu Referenzübersetzung:

$$\text{WER} = \frac{\text{min(substitutionen + einfügungen + löschungen)}}{\text{ref-länge}}$$

Word Error Rate (WER): Beispiel

Hypothese:
Israeli officials responsibility of airport safety

Referenz:
Israeli officials are responsible for airport security

$$\text{WER} = \frac{\text{min(substitutionen + einfügungen + löschungen)}}{\text{ref-länge}} =$$

Word Error Rate (WER): Beispiel

Hypothese:
Israeli officials responsibility of airport safety

Referenz:
Israeli officials are responsible for airport security

$$\text{WER} = \frac{\text{min(substitutionen + einfügungen + löschungen)}}{\text{ref-länge}} = \frac{4}{7} = 0.571 = 57.1\,\%$$

Word Error Rate (WER): Problem

Hypothese:
This airport's security is the responsibility of the Israeli security officials

Referenz:
Israeli officials are responsible for airport security

Word Error Rate (WER): Problem

Hypothese:
This airport's security is the responsibility of the Israeli security officials

Referenz:
Israeli officials are responsible for airport security

WER $>100\,\%$

Word Error Rate (WER): Problem

Hypothese:
This airport's security is the responsibility of the Israeli security officials

Referenz:
Israeli officials are responsible for airport security

WER $>100\,\%$ $\rightarrow$ Strikte Einhaltung der Wortreihenfolge ist zu harsch

Translation Error Rate[1] (TER)

TER (Snover et al., 2006) ist WER mit einem Zusatz: Jede Verschiebung einer Mehrwertsequenz (phrasal shift) zählt als eine Editieroperation.

[1] Wird auch als Translation Edit Rate bezeichnet.

Bilingual Evaluation Understudy (BLEU)

BLEU (Papineni et al., 2002) ist die wahrscheinlich populärste automatische Evaluationsmetrik für Übersetzungsqualität. Die Leitideen sind:

- Berechnung von n-Gramm-Überlappungen der Hypothese mit mehreren Referenzübersetzungen[1]
- Kein Recall; kompensiert durch «**Brevity Penalty**»
- Finaler Wert ist das gewichtete geometrische Mittel der **n-Gramm-Präzision** (i.d.R. mit n=1,2,3,4).
- Berechnung auf Testkorpus- statt Satzebene, da n-Gramm-Präzision für höhere Ordnungen (z.B. n=4) oft 0 ist.

[1] In der Praxis wird oft nur eine Referenzübersetzung verwendet.

BLEU: Brevity Penalty

$$\text{BP} = \min\left(1.0, \exp\left(1 - \frac{\text{ref-länge}}{\text{hyp-länge}}\right)\right)$$

- Bestrafung, wenn Hypothese kürzer als Referenz ist
- Bei mehreren Referenzen: Länge der Referenz, die am nähsten bei der Länge der Hypothese liegt (s. Koehn, 2010, S. 227)

BLEU: n-Gramm-Präzision

$$\mathrm{P} = \left(\prod_{n=1}^{N} \lambda_n p_n \right)^{\frac{1}{N}}$$

- N: höchste n-Gramm-Ordnung (i.d.R. 4)
- n: n-Gramm-Präzision der Ordnung n
- λ_n: Gewicht der n-Gramm-Präzision der Ordnung n (i.d.R. 1.0)

BLEU

$$\begin{aligned}\text{BLEU} &= \text{BP} \cdot \text{P} \\ &= \min\left(1.0, \exp\left(1 - \frac{\text{ref-länge}}{\text{hyp-länge}}\right)\right) \cdot \left(\prod_{n=1}^{N} \lambda_n p_n\right)^{\frac{1}{N}}\end{aligned}$$

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme:

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible)

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme:

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible)

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2 = 4/5$

3-Gramme:

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2 = 4/5$

3-Gramme: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible)

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2 = 4/5$

3-Gramme: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3 = 2/4$

4-Gramme:

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2 = 4/5$

3-Gramme: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3 = 2/4$

4-Gramme: ~~(airport security Israeli officials)~~ ~~(security Israeli officials are)~~ (Israeli officials are responsible)

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2 = 4/5$

3-Gramme: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3 = 2/4$

4-Gramme: ~~(airport security Israeli officials)~~ ~~(security Israeli officials are)~~ (Israeli officials are responsible) $\rightarrow p_4 = 1/3$

Brevity Penalty:

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

1-Gramme: (airport) (security) (Israeli) (officials) (are) (responsible) $\rightarrow p_1 = 6/6$

2-Gramme: (airport security) ~~(security Israeli)~~ (Israeli officials) (officials are) (are responsible) $\rightarrow p_2 = 4/5$

3-Gramme: ~~(airport security Israeli)~~ ~~(security Israeli officials)~~ (Israeli officials are) (officials are responsible) $\rightarrow p_3 = 2/4$

4-Gramme: ~~(airport security Israeli officials)~~ ~~(security Israeli officials are)~~ (Israeli officials are responsible) $\rightarrow p_4 = 1/3$

Brevity Penalty: $\min\left(1.0, \exp\left(1 - \frac{7}{6}\right)\right) = 0.846$

BLEU: Beispiel

Hypothese:
airport security Israeli officials are responsible

Referenz:
Israeli officials are responsible for airport security

$$
\begin{aligned}
\text{BLEU} &= BP \cdot (p_1 \cdot p_2 \cdot p_3 \cdot p_4)^{\frac{1}{4}} \\
&= 0.846 \cdot \left(\frac{6}{6} \cdot \frac{4}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \right)^{\frac{1}{4}} \\
&= 0.511 \\
&(= \text{wird oft als 51.1, also in Prozent, angebgeben.})
\end{aligned}
$$

BLEU: Mehrere Referenzen

Sind mehrere Referenzübersetzungen vorhanden, gilt:

- Ein n-Gramm zählt als abgedeckt, wenn es in *irgendeiner* Referenz vorkommt. (Achtung: Clipping!)
- Für die Brevity Penalty dient
  - die Länge derjenigen Referenzübersetzung, die der Länge der Hypothese am nähsten kommt;
  - der kleinere Wert, wenn zwei Referenzlängen (z.B. 9, 11) den gleichen Abstand zur Hypothesenlänge (z.B. 10) haben.

BLEU: Clipping

Hypothese:
are are are are are are are

Referenz:
Israeli officials are responsible for airport security

Jedes n-Gramm zählt nur so oft als abgedeckt, wie es innerhalb einer Referenzüberetzung vorkommt.

BLEU: Clipping

Hypothese:
are are are are are are are

Referenz:
Israeli officials are responsible for airport security

Jedes n-Gramm zählt nur so oft als abgedeckt, wie es innerhalb einer Referenzüberetzung vorkommt.

$\rightarrow$ 1-Gramm-Präzision ist 1/7, nicht 7/7!

BLEU: Clipping – Beispiel

Hypothese:
the the the the the the the

Referenz 1:
the cat is on the mat

Referenz 2:
there is a cat on the mat

1-Gramm-Präzision p_1 =
2-Gramm-Präzision p_2 =

BLEU: Clipping – Beispiel

Hypothese:
the the the the the the the

Referenz 1:
the cat is on the mat

Referenz 2:
there is a cat on the mat

1-Gramm-Präzision p_1 = 2/7
2-Gramm-Präzision p_2 =

BLEU: Clipping – Beispiel

Hypothese:
the the the the the the the

Referenz 1:
the cat is on the mat

Referenz 2:
there is a cat on the mat

1-Gramm-Präzision p_1 = 2/7
2-Gramm-Präzision p_2 = 0/7

BLEU: Probleme

- **Ignoriert Relevanz verschiedener Wörter**

  Gewisse Wörter sind wichtiger für eine Übersetzung als andere; in BLEU werden alle gleich gewichtet.

  - Beispiel: Unübersetzte Wörter
  - Referenz: «gave it to Trump»
  - Hypothese «gave it at Trump» schneidet schlechter ab als «gave it to rhododendron»

- **BLEU-Wert an und für sich ist bedeutungslos**

  Wert hängt von vielen Faktoren wie Anzahl Referenzübersetzungen, Sprache und Domäne ab – und von Vorverarbeitungsschritten wie z.B. der Tokenisierung.

- **Mit steigender Verbesserung von MT eignet sich BLEU immer weniger**

  Ist BLEU noch gut genug für Neuronale Maschinelle Übersetzung (NMT)?

---

s. auch Callison-Burch et al., 2006

METEOR

METEOR (Banerjee und Lavie, 2005) ist eine beliebte Alternativ- oder Komplementärmetrik zu BLEU.

- Leitidee: Recall ist wichtiger als Präzision um sicherzustellen, dass die komplette Bedeutung abgedeckt ist.
- Alignierung von Wörtern in Hypothese und Referenzübersetzung(en)
- Dreistufiges Matching:
  - **Oberflächenform**; ansonsten
  - **Stamm** (mittels Stemming) mit Abzug; ansonsten
  - **Semantische Klasse** (mittels Wordnet) mit Abzug; ansonsten
  - keine Übereinstimmung

METEOR: Probleme

- viele Parameter (z.B. Gewichte für Stamm- und Synonym-Matches)
- kompliziertere Berechnung als BLEU
- Sprachabhängig: benötigt Stemmer und Synonyme
- rechenintensiv (Alignment, Stemming, Synonym-Lookup, etc.)

Themen

1. Einführung

2. Manuelle Evaluation

3. Automatische Evaluation

4. Zusammenfassung

Überblick: Manuelle Evaluation

Methode	**Merkmale**	**Probleme**
Absolute Evaluation: Adäquatheit und Flüssigkeit	• Likert-Skala (i.d.R. 5 Punkte)	• schwer interpretierbar • schlecht reproduzierbar
Relative Evaluation: Ranking	• Rangierung von 2 oder mehr Systemen • einfach interpretierbar	• nur Ordnung, nicht Ausmass des Unterschieds

Überblick: Automatische Evaluation

Methode		Merkmale	Probleme
Precision, Recall, F-Measure	↑	• einfache Berechnung	• Wortreihenfolge wird ignoriert
WER (Word Error Rate)	↓	• einfache Berechnung (minimale Editierdistanz)	• Strikte Einhaltung der Wortreihenfolge zu harsch
TER (Translation Edit Rate)	↓	• wie WER • zusätzliche Editieroperation: Blockverschiebung	• s. BLEU
BLEU	↑	• Fokus: Precision • sehr weit verbreitet	• alle Wörter gleich gewichtet • Wert an und für sich nicht interpretierbar
METEOR	↑	• Fokus: Recall • weit verbreitet • Wortalignierung, Stemming, Synonymie	• sprachabhängig • komplizierte, rechenintensive Berechnung • Wert an und für sich nicht interpretierbar

↑ = höherer Wert ist besser; ↓ = tieferer Wert ist besser

Literatur I

Banerjee, Satanjeev und Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan, Seiten 65–72.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, und Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*. Sofia, Bulgaria, Seiten 1–44.

Callison-Burch, Chris, Miles Osborne, und Philipp Koehn. 2006. Re-evaluation the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy, Seiten 249–256.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.

Literatur II

Koehn, Philipp und Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation (WMT)*. New York, NY, USA, Seiten 102–121.

Papineni, Kishore, Salim Roukos, Todd Ward, und Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*. Philadelphia, PA, USA, Seiten 311–318.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, und John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*.