

Fortgeschrittene Themen
der statistische maschinelle Übersetzung
(Advanced SMT)
Evaluierung

Miriam Kaeshammer
Heinrich-Heine-Universität Düsseldorf

Folien angepasst von T. Schoenemann

Wintersemester 2014/15

Überblick

Frage: **Wie bewerten wir die Qualität eines (automatisch) übersetzten Satzes/Textes?**

Anwendungen:

- bei der Publikation eines Papers
- regelmäßige Evaluierungen/Wettbewerbe, bei denen Gruppen einreichen können
- auch für Firmen intern, um die Weiterentwicklung ihrer Systeme steuern zu können
- zur Systemoptimierung (Tuning)

Evaluierungsmöglichkeiten:

- Manuelle Auswertung
- Automatische Auswertung: vgl. mit *Referenzübersetzung(en)*
- Downstream-Evaluierung: z.B. Informationsextraktion aus einem fremdsprachigen Text

Komplikationen

- Es gibt nicht *die eine richtige* Ausgabe.
 - Möglichst mehrere Referenzübersetzungen
- praktisch unmöglich, *alle* akzeptablen Übersetzungen zu erfassen (z.B. Wortordnung oft sehr variabel)
- Auswertung durch Menschen ist sehr teuer und dauert zu lange.

Manuelle Übersetzungen eines chinesischen Satzes von unterschiedlichen Übersetzern:

- Israeli officials are responsible for airport security.
- Israel is in charge of the security of this airport.
- The security work for this airport is the responsibility of the Israel government.
- Israeli side was in charge of the security of this airport.
- Israel is responsible for the airport's security.
- ...

Manuelle Evaluierung

Mehrere menschliche Experten bewerten die gleiche Menge von Hypothesen

Ideal: **bilinguale Bewerter**

– jedoch schwer zu kriegen

In der Praxis: meist **monolinguale Bewerter, die zu Referenzübersetzungen vergleichen.**

Manuelle Evaluierung ist sehr **subjektiv.**

- z.B.: manche Hypothesen ergeben zunächst keinen Sinn, werden aber klar, wenn man die Referenz oder den Eingabesatz liest.
- Teilgrund: Sätze ohne Kontext sind generell schwer verständlich.

Einzelkriterien

Fluency und Adequacy, jeweils Skala 1-5

Adequacy:

5 - gesamter Inhalt wiedergegeben

4 - der Großteil des Inhalts ist vorhanden

3 - ein ordentlicher Anteil des Inhalts ist vorhanden

2 - nur wenig Inhalt vorhanden

1 - fast gar nichts vorhanden

Fluency:

5 - fehlerlos

4 - gut

3 - Niveau eines Fremdsprachlers

2 - stark verzerrt

1 - nicht verständlich

Beispiel

Dt.: aber ich will nicht nach hause gehen !

Referenz: but i don't want to go home !

Hypothese: i want not go home but !

Adequacy: ca. 4

Fluency: ca. 2

Gründe für automatische Metriken

Evaluierung ist für **Tuning** extrem wichtig

Tuning = Anpassung von Gewichtungsparemtern, sodass die Übersetzungsqualität optimiert wird

$$\max_{\{\lambda_i\}} \text{EVAL} \left(\left\{ \operatorname{argmax}_{\vec{e}} \sum_i \lambda_i h_i(\vec{e}, \vec{f}_s) \mid s = 1, \dots, S \right\} \right)$$

Dazu notwendig: Evaluierungsscores für $500 < S < 5000$

Hypothesen (Entwicklungsdatensatz, *development data*)

⇒ Scores müssen automatisch berechenbar sein.

Außerdem: menschliche Bewerter müssen bezahlt werden, automatische Evaluierung verursacht praktisch keine Kosten.

Gewünscht: Metrik, die gut mit menschlichen Scores korreliert.

Basis: Vergleich zwischen Hypothese **h** und Referenz **r**

***n*-gram-basierte Metriken**

Basis: Anzahl der korrekt vorhandenen *n*-gramme für unterschiedliche *n*.

Für Hypothese $\mathbf{h} = h_1^H$, Referenz $\mathbf{r} = r_1^R$:

***n*-gram Precision:**

$$\frac{\#n\text{-grams present in } \mathbf{h} \text{ and } \mathbf{r}}{\#n\text{-grams present in } \mathbf{h}}$$

***n*-gram Recall:**

$$\frac{\#n\text{-grams present in } \mathbf{h} \text{ and } \mathbf{r}}{\#n\text{-grams present in } \mathbf{r}}$$

F-measure: Kombination aus Precision und Recall (aber selten benutzt)

Beispiel

Referenz Israeli officials are responsible for airport security
 Hypothese A Israeli officials ~~responsibility~~ of airport ~~safety~~
 Hypothese B airport security Israeli officials are responsible

Für Hypothese A:

1-gram Precision: $\frac{3}{6}$	2-gram Precision: $\frac{1}{5}$	3-gram Precision: $\frac{0}{4}$
1-gram Recall: $\frac{3}{7}$	2-gram Recall: $\frac{1}{6}$	3-gram Recall: $\frac{0}{5}$

Für Hypothese B:

1-gram Precision: $\frac{6}{6}$	2-gram Precision: $\frac{4}{5}$	3-gram Precision: $\frac{2}{4}$
1-gram Recall: $\frac{6}{7}$	2-gram Recall: $\frac{4}{6}$	3-gram Recall: $\frac{2}{5}$

BLEU: A bilingual evaluation understudy

n-gram Precisions für unterschiedliche *n* + Längenstrafterm

$$\text{BLEU-}n : \min \left(1, \frac{H}{R} \right) \exp \left(\sum_{k=1}^n \lambda_k \log (k\text{-precision}) \right)$$

Üblich: $\lambda_k = 1$, BLEU-4

Beachte: höhere Werte = bessere Übersetzung

BLEU: Beispiel

Referenz Israeli officials are responsible for airport security

Hypothese A Israeli officials ~~responsibility~~ of airport ~~safety~~

Hypothese B airport security Israeli officials are responsible

	n	1	2	3	4
Hypo A	n -gram Prec.	$\frac{3}{6}$	$\frac{1}{5}$	0	0
	BLEU- n	$\frac{6}{7} \cdot \frac{3}{6} \approx \mathbf{0,42}$	$\frac{6}{7} \cdot \frac{3}{6} \cdot \frac{1}{5} \approx \mathbf{0,09}$	0	0
Hypo B	n -gram Prec.	$\frac{6}{6}$	$\frac{4}{5}$	$\frac{2}{4}$	$\frac{1}{3}$
	BLEU- n	0,86	0,69	0,34	0,11

Problem: Score ist 0 sobald eine n -gram Precision 0 ist.

→ Auswertung/Normalisierung auf Korpus-Level, nicht für jeden Satz einzeln

→ Anpassung für mehrere Referenzübersetzungen

Kritik an BLEU

- Wörter sind entweder völlig falsch oder völlig richtig
- Aber: **responsibility** und **responsible** sind ähnlich
⇒ Inhalt des Satzes teilweise vorhanden

METEOR:

Einbeziehung von Ähnlichkeiten/Synonymen durch Stemming und WordNet

Probleme von METEOR:

- Viele Parameter beteiligt (wie setzen?)
- WordNet etc. sind work-in-progress und nicht für alle Sprachen vorhanden
- Schwierig, ein Masterprogramm für alle Sprachen zu erstellen
Insbesondere: WordNet belegt einigen Speicherplatz!

Edit-basierte Metriken

Prinzip: Die Referenzübersetzung wird durch elementare Operationen schrittweise in die gegebene Hypothese umgeformt.

Word Error Rate (WER): elementare Operationen:

- Ersetzen eines Wortes durch ein anderes
- Einfügen eines Wortes
- Löschen eines Wortes

WER für:

- einen Satz: minimale Anzahl von Operationen, um die Referenz in die Hypothese zu transformieren, normalisiert durch die Referenzlänge
- eine Menge von Sätzen: Mittelwert der Einzelsatz-WERs

Beachte: kleinere Werte = bessere Übersetzungen

Bestimmung von WER (Satzebene)

Bestimmung eines monotonen Alignments (genannt *Levenshtein-Alignment*) zwischen Hypothese und Referenz:

- Matching von identischen Wörtern: Score unverändert
- Matching von nicht-identischen Wörtern: erhöht Score um 1.
- Referenzwort ohne Alignment (= löschen): erhöht Score um 1.
- Hypothesenwort ohne Alignment (= einfügen): erhöht Score um 1.

→ **WER: Levenshtein-Alignment mit minimalem Score**

Minimales Levenshtein-Alignment

Dynamische Programmierung:

Tabelle $Q(i, j)$ mit $0 \leq i \leq R, 0 \leq j \leq H$.

Basisfall: $Q(0, 0) = 0$

Aufbaufall ($i \geq 1$ oder $j \geq 1$):

$$Q(i, j) = \min\{\begin{array}{l} Q(i-1, j-1) \quad \text{falls } r_i = h_j, \quad \% \text{ match} \\ Q(i-1, j-1) + 1, \quad \% \text{ substitute} \\ Q(i-1, j) + 1, \quad \% \text{ delete} \\ Q(i, j-1) + 1 \quad \% \text{ insert} \end{array}\}$$

(wobei Scores für $i = -1$ oder $j = -1$ als ∞ definiert sind)

Levenshtein Alignment: Beispiel

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

Berechnung WER

$$\text{WER}_{r,h} = \frac{Q(R,H)}{R}$$

Falls gewünscht, lässt sich das entsprechende Alignment durch Traceback ermitteln.

Verbesserungen von WER

Problem von WER:

- Umordnungen nicht explizit modelliert, somit stark bestraft.

Referenz 1 Israeli officials are responsible for airport security

Referenz 2 This airport's security is the responsibility of the
Israeli security officicals

Integration von Umordnungen: → Translation Edit Rate (TER)

Basis: **Block-Moves** zusätzlich zu den normalen Edit Operationen

Kosten: 1

Diskussion (1)

- BLEU-4 momentan akzeptierter Standard (auch beliebt: TER)
- BLEU-Scores korrelieren mit manuellen Scores
(Arabisch-Englisch, NIST 2002)

Jedoch:

- Für BLEU sind alle Wörter gleich relevant: Verneinung, Inhaltswörter vs. Artikel, Satzzeichen?
- Niemand weiß, was 0,34 BLEU heißt.
- BLEU arbeitet sehr lokal → Verdacht, dass BLEU phrasenbasierte Systeme gegenüber baumbasierten unfair bevorteilt.

Diskussion (2)

Experimente:

- Regelbasiertes System vs. statistisches:
Statistisches bekam höhere BLEU-Scores, aber viel niedrigere manuelle Bewertungen
- (monolingual) manuell verbesserte Übersetzungen bekamen nur leicht bessere BLEU-Scores, aber viel bessere manuelle Bewertungen

Ähnliche Argumente lassen sich auch für die anderen automatischen Metriken finden.

Andere Evaluierungskriterien

Neben Qualität:

- Schnelligkeit
- System-/Modellgröße (→ Server vs. Smartphone)
- Integration in eine Anwendungsumgebung
- Anpassung (andere Domäne, Kundenwünsche etc.)