Statistische Maschinelle Übersetzung

Teil V - Evaluierung

Thomas Schoenemann
Heinrich-Heine-Universität Düsseldorf

Sommersemester 2012

Überblick

Frage in diesem Teil der Vorlesung: **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.

Komplikationen:
– es gibt mehrere richtige Übersetzungen
– praktisch unmöglich, *alle* akzeptablen Übersetzungen zu erfassen
(z.B. Wortordnung oft sehr variabel)
– Auswertung durch Menschen ist sehr teuer und dauert zu lange.
Bei Wettbewerben teils Manipulationen vermutet.

Mehrdeutigkeit

Manuelle Übersetzungen eines chinesischen Satzes von 10 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.`
- `Israel is responsible for safety work at this airport.`
- `Israel presides over the security of this airport.`
- `Israel took charge of the airport security.`
- `The safety of this airport is taken charge of by Israel.`
- `This airport's security is the responsibility of the Israeli security officials.`

Mehrdeutigkeit und Evaluierungen

Wegen Mehrdeutigkeiten:
– möglichst mehrere Referenzübersetzungen für die Evaluierung verwenden
– jedoch: meist nur eine vorhanden

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.

Manuelle Evaluierung: Einzelkriterien

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 !`

Fluency ca. 2

Adequacy: ca. 4

Ausgleich zwischen Bewertern

Ziel bei mehreren Bewertern: jedem Satz einen einzigen Score zuordnen

Komplikation:
unterschiedliche Bewerter sind unterschiedlich großzügig

Daher: Vorverarbeitung der Scores der Einzelbewerter:
– zunächst Subtraktion des Mittelwertes
– evtl. Normalisierung bzgl. Varianzen
– schließlich: Addition des Mittelwertes der Mittelwerte aller Bewerter

$\Rightarrow$ Scores jetzt besser vergleichbar, mitteln der Einzelscores

Gründe für automatische Metriken

Evaluierung ist für Tuning extrem wichtig

Tuning = Anpassung von Gewichtungsparametern, sodass die Übersetzungsqualität optimiert wird

Dazu notwendig: Evaluierungsscores für $500 - 5000$ Hypothesen, innerhalb von Sekunden - für jeden getesteten Parametersatz

$\Rightarrow$ 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 und Referenz

Im Folgenden: Hypothese $\mathbf{h}$, Referenz $\mathbf{r}$

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:

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

n-gram Recall:

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

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

Precision und Recall: 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`

Intern: Erweiterung um $<s>$ und $</s>$

Für Hypothese B:

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

1-gram Recall: $\frac{6}{7}$ 2-gram Recall: $\frac{4}{8}$

Für Hypothese A:

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

1-gram Recall: $\frac{3}{7}$ 2-gram Recall: $\frac{2}{8}$

BLEU : A bilingual evaluation understudy

BLEU: Längenstrafterm + n-gram Precisions für unterschiedliche n

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

Üblich: $\lambda_k = 1$, BLEU-4

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

Abhilfe: Auswertung/Normalisierung auf Corpus-Level, nicht für jeden Satz einzeln

Beachte: Accuracy-Measure, höhere Werte = bessere Übersetzung

METEOR

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 durch Stemming und WordNet bzw. anderssprachige Equivalente

(`house` ist ähnlicher zu `home` als zu `mouse`)

Probleme von METEOR:

– Viele Parameter beteiligt (wie setzen?)

– WordNet etc. sind work-in-progress

– Schwierig, ein Masterprogramm für alle Sprachen zu erstellen

Insbesondere: WordNet belegt einigen Speicherplatz!

– Komplikationen: Zeichensatz, Groß/Kleinschreibung etc.

Edit-basierte Metriken

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

Word Error Rate: elementare Operationen:
– Ersetzen eines Wortes durch ein anderes
– Einfügen eines Wortes
– Löschen eines Wortes

Word Error Rate ... :
– eines Satzes: min. Anzahl von Operationen, um die Referenz in die Hypothese zu transformieren, normalisiert durch die Referenzlänge
– einer Menge von Sätzen: Mittelwert der Einzelsatz-WERs

Beachte: WER ist Fehlermaß, kleinere Werte = bessere Übersetzungen

Bestimmung von WER

Für Bestimmung der 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.

Für WER: Bestimmung des Levenshtein-Alignments mit minimalem Score

Minimales Levenshtein-Alignment

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

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

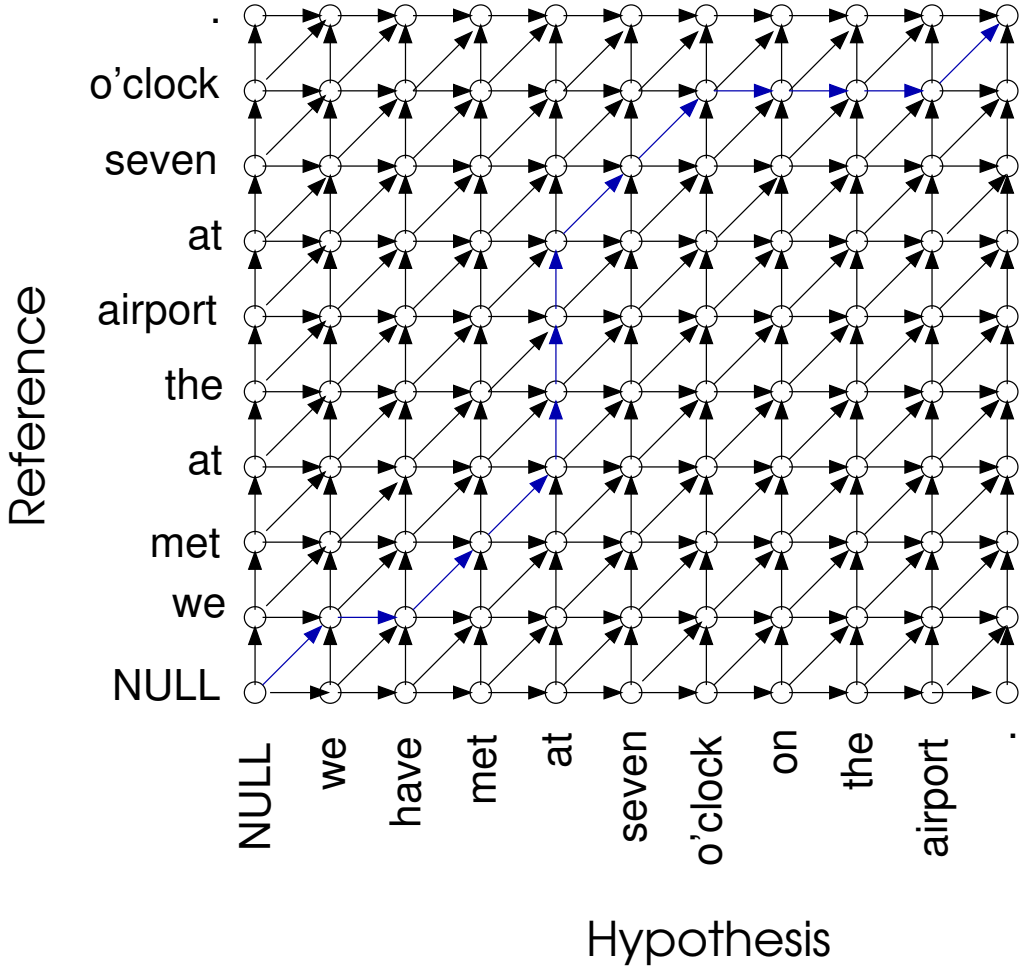
Aufbaufall ($i \geq 1$ oder $j \geq 1$):

$$\begin{aligned} Q(i,j) = \min\{ & \delta(h_i, r_j) + Q(i-1, j-1), \quad \texttt{\% match} \\ & 1 + Q(i, j-1), \quad \texttt{\% insert} \\ & 1 + Q(i-1, j)\} \quad \texttt{\% delete} \end{aligned}$$

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

$\frac{Q(H,R)}{R}$ gibt die WER, das entsprechende Alignment lässt sich (falls gewünscht) durch Traceback ermitteln

Levenshtein Alignment: Zustandsraum



reproduziert von [Leusch et al. 2006]

Verbesserungen von WER

Problem von WER:

– Umordnungen nicht explizit modelliert, somit stark bestraft.

Integration von Umordnungen: → Translation Edit Rate

– Basis: Block-Moves zusätzlich zu den normalen Edit Operationen
  Kosten: 1

– mit beliebigen Umordnungen der Hypothese: NP-hart
  (Lopresti und Tomkins, 1997 ; Shapira und Storer, 2002)

– [Snover et al. 2005] : heuristische Suche

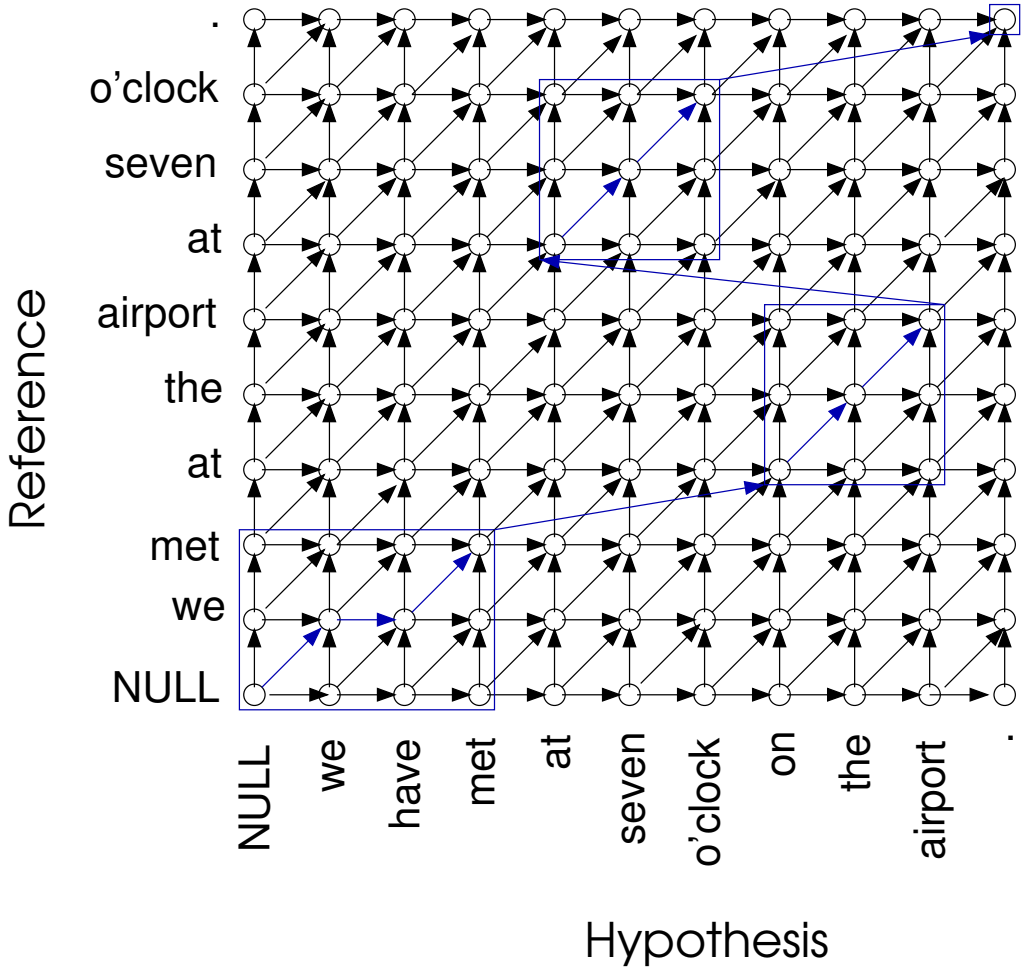
– [Leusch et al. 2003] : Einschränkung der Umordnungen
  (polynomiell, aber hoher Exponent)

– [Leusch et al. 2006] : modifiziertes Kriterium

Problem also:

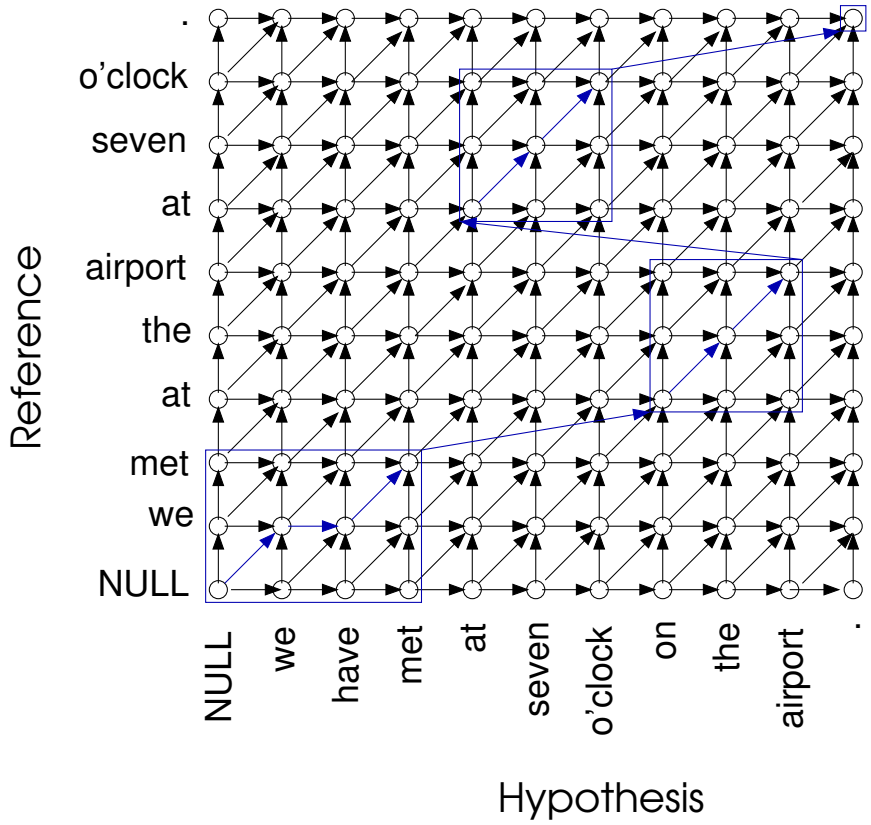
TER sinnvolles Fehlermaß, aber leider nicht (exakt) berechenbar

Block Moves



reproduziert von [Leusch et al. 2006]

Block Moves: Erlaubte Pfade



Ein erlaubter Pfad

– deckt jedes Hypothesenwort genau einmal ab

– deckt jedes Referenzwort genau einmal ab

Diskussion

BLEU-4 momentan akzeptierter Standard, (auch beliebt: TER)

Jedoch:

- Experiment: Regelbasiertes System vs. Statistisches:
  Stat. bekam höhere BLEU-Scores, aber viel niedrigere manuelle Bewertungen
- Anderes Experiment: (monolingual) manuell verbesserte Übersetzungen bekamen nur leicht bessere BLEU-Scores, aber viel bessere manuelle Bewertungen
- Verdacht, dass BLEU phrasenbasierte Systeme gegenüber baumbasierten unfair bevorteilt.

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