Evaluation von Wortalinierung für MÜ (Fraser&Marcu 2007)

Sebastian Pado

Überblick

- Evaluation gegen Goldstandard vs. task-basierte Evaluation
  - Maschinelle Übersetzung
- Qualitätsmaße: F-Score vs. AER
  - Mögliche vs. sichere Links
- Task-basierte Evaluation (Maschinelle Übersetzung)

Zwei Möglichkeiten der Evaluation:

- Gegen ein annotiertes Korpus
  - "in vitro"-/Goldstandard-basierte Evaluation
- Oder in einem System / an einer konkreten Aufgabe
  - "in vivo" / task-basierte Evaluation
  - Alinierung ist eine Komponente in einem komplexen System

- Oft wird die Evaluation gegen ein annotiertes Korpus als 'manuelle' Evaluation bezeichnet
  - Evaluation in eine System kann aber automatisch oder manuell erfolgen…

- MÜ: großes Interesse an taskbasierter Evaluation
  - Alinierungen dienen zur Extraktion des Übersetzungslexikons
  - Zu optimierendes Endprodukt: Übersetzungsqualität

Vor- und Nachteile

- Manuelle Evaluation
  - ✓Generalisierbare Bewertung (idealerweise)
  - ✗ Keine Aussage für konkrete Aufgabe
  - ✗ Teuer (zumindest im Vergleich zu vollautomatischer Evaluation)

- Task-basierte Evaluation
  - ✓ Spezifische Bewertung der Qualität eines konkreten Systems
  - ✗ Spezifische Bewertung der Qualität eines konkreten Systems
  - ✗ Bewertet nicht nur Wortalinierung, sondern auch die Qualität aller anderer Komponenten

Der heilige Gral der Evaluation

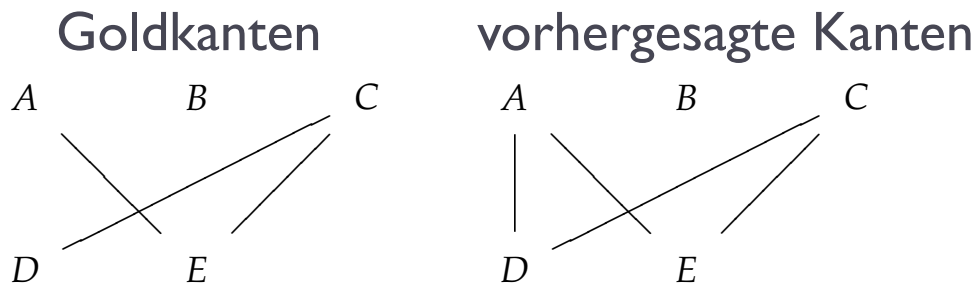
- Idealzustand: Manuelle Evaluation ist **prädiktiv** für die taskbasierten Evaluationsergebnisse
  - Schritt 1: Ein Korpus manuell annotieren
  - Schritt 2: Wortalinierung optimieren
  - Schritt 3: Dann alle anderen Komponenten optimieren

- Erfahrung in Maschineller Übersetzung 2000 bis 2005:
  - Neue Modelle: bessere Wortalinierung
    - laut manueller Evaluation
  - ...aber tatsächlich kaum bessere Übersetzungsqualität
    - laut taskbasierter Evaluation

- **Wo liegt das Problem? -- Fraser & Marcu 2007**

Manuelle Evaluation: Qualitätsmaße

- Wortalinierungen können als Kanten (“links”) in einem Alinierungsgraphen verstanden werden
- Evaluation: vergleiche vorhergesagte Kanten (A) und Goldkanten (G)

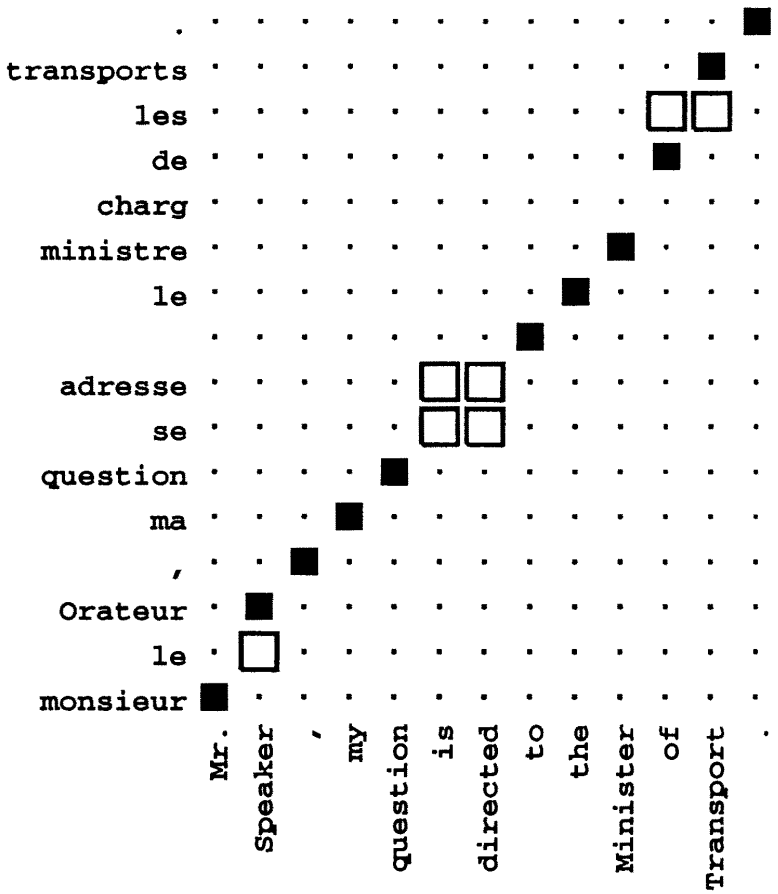


Präzision: Pr = |G ∩ A| / |A| = 3/4 = 0.75

Recall: R= |G ∩ A| / |G| = #3/3 = 1

F-Score: 2*Pr*R/(Pr+R) = 0.86

Mögliche und sichere Kanten



- Wörtliche Entsprechungen: "sure links" (S)
- Entsprechungen im weiteren Sinne: "possible links" (P)

- $G = P,\ S \subseteq P$
- $Pr = |P \cap A| / |A|$
  - possible links dürfen vorausgesagt werden
- $R = |S \cap A| / |S|$
  - fehlende possible links zählen nicht gegen den Recall

Alignment Error Rate

- Das von 2000 bis 2005 am weitesten verwendete Evaluationsmaß für Wortalinierungen

$$AER = 1 - \frac{|P \cap A| + |S \cap A|}{|A| + |S|}$$

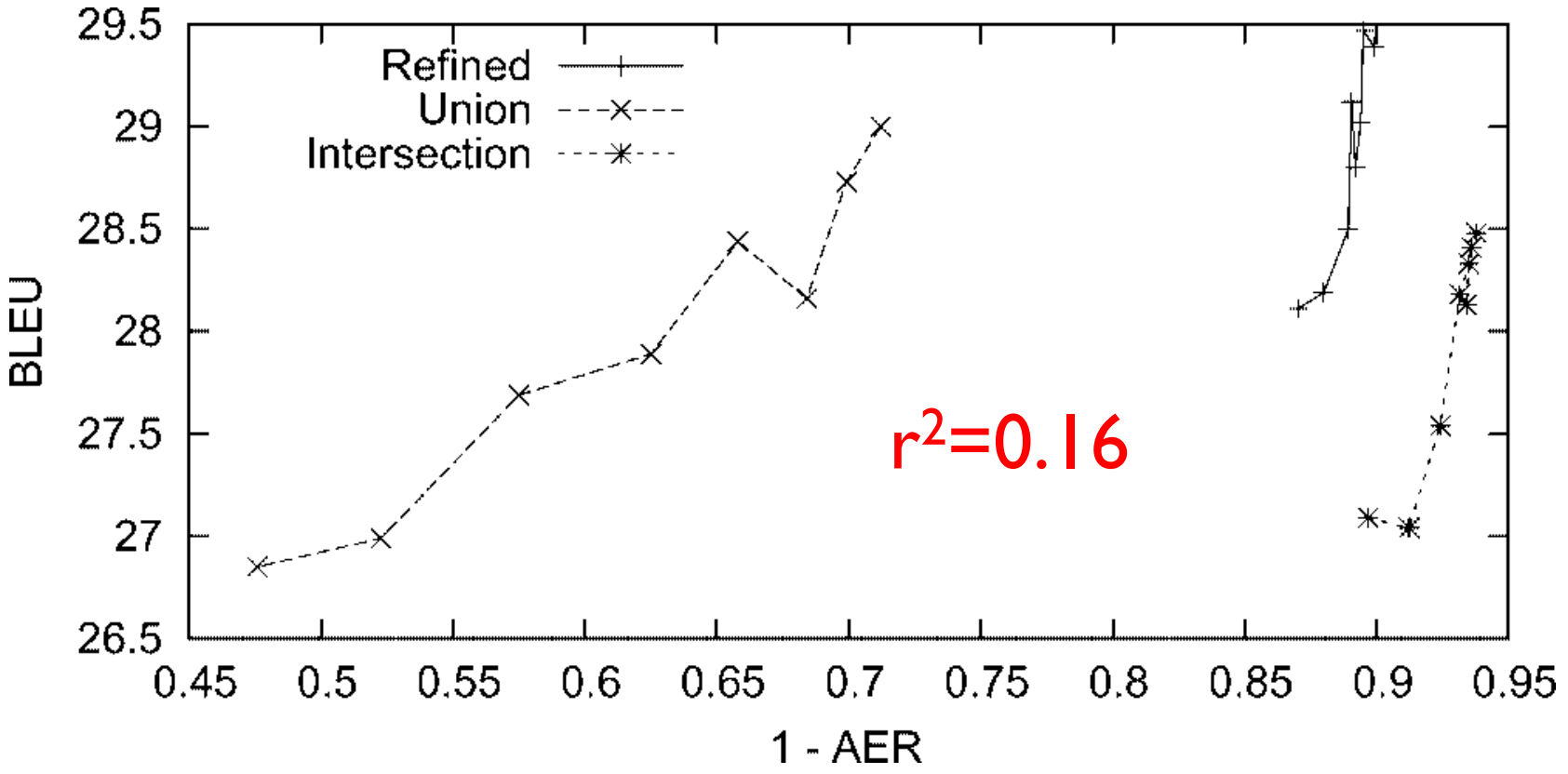
("motiviert durch F-Score")

- Niedrige Werte: gute Alinierung
  - Analog zu Word Error Rate
  - Ich verwende (wie F&M) i.A. 1-AER (hoch=gut)

F&M: Analyse der AER-Ergebnisse

- Kontrollierte Variation der Alinierungsqualität
  - Verkleinerung des Trainingskorpus
- Vergleich mit Übersetzungsqualität
  - Übersetzungsqualität approximiert durch BLEU (Papineni et al. 2001)
    - n-gram-basiertes Maß
    - Überlappung mit einer oder mehreren Referenzübersetzung(en)
- Operationalisierung des Vergleichs: Korrelation
  - Linearer Korrelationskoeffizient (Pearson's r): {-1…0…1}
  - Einfacher **Vorhersagemechanismus**
  - r^2: Anteil der **erklärten Varianz**

Alinierung Französisch-Englisch



Wo liegt das Problem?

- In der Definition von AER!
  - AER entspricht angeblich F-Score mit possible und sure links
  - Das **stimmt aber nicht**

$$1 - AER = \frac{|P \cap A| + |S \cap A|}{|A| + |S|}$$

- F-Score = 2*Pr*R / (Pr+R) = 1/ (1/Pr + 1/R)

$$= \frac{|P \cap A| * |S \cap A| + |P \cap A| * |S \cap A|}{|A||S \cap A| + |S||P \cap A|}$$

F-Score vs. AER

- Warum ist dieser Unterschied relevant?
  - Was ist an F-Score so besonderes?

- Der normale F-Score **balanciert** Precision und Recall
  - Wenn P = R, dann F=P=R
  - Wenn R < P, dann F < R< P (und umgekehrt)
- **Das ist bei AER nicht der Fall**

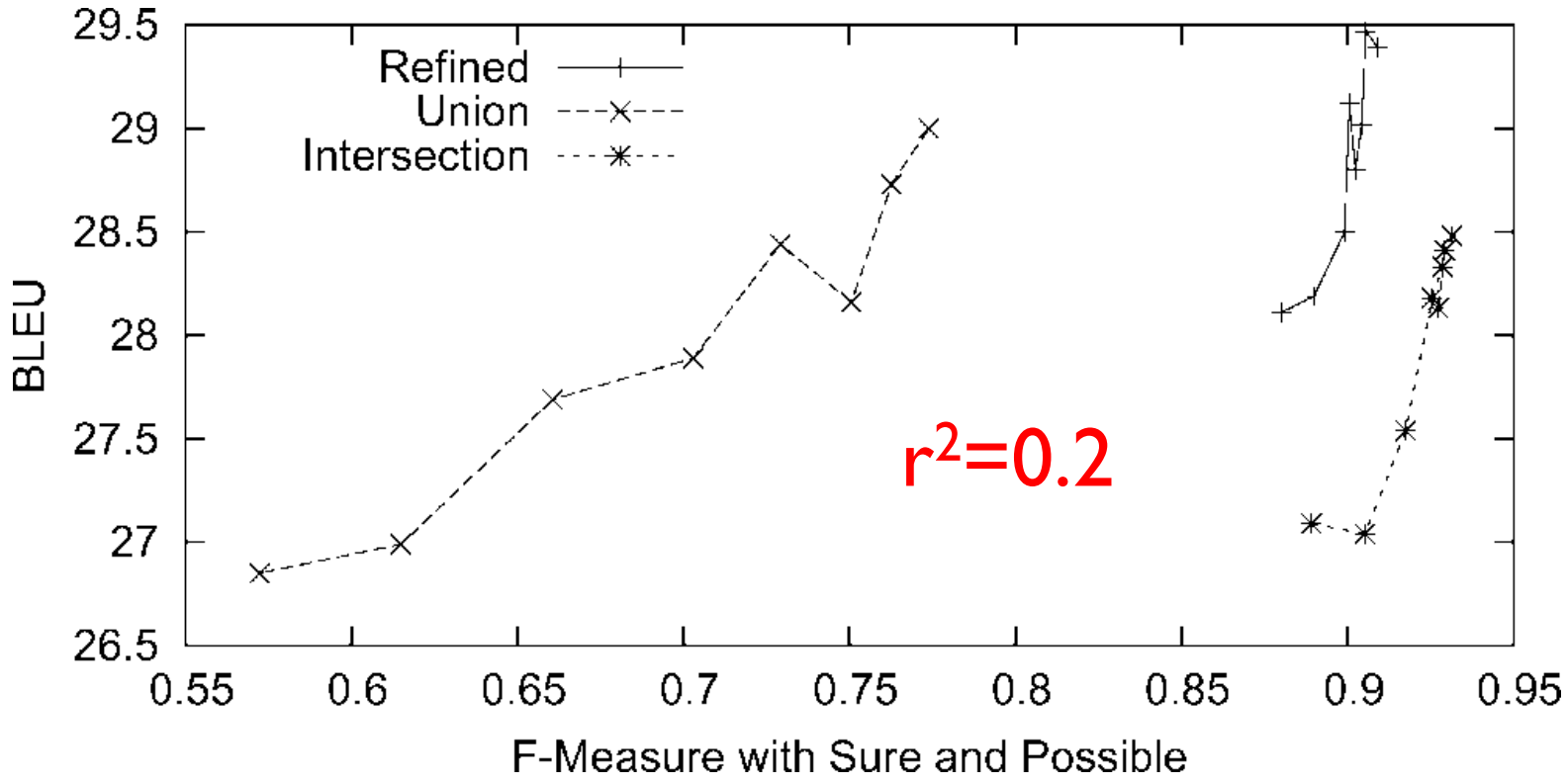
Gegenbeispiel

- Gold-Standard "sure": A, B
- Gold-Standard "possible": A,B,C, D
- Alinierung 1: A, B, K, L
- Alinierung 2: A, C, D, L

- Alignment 1: Prec 0.5, Rec. 0.5, F-Score 0.5, AER 0.5
- Alignment 2: Prec 0.75, Rec 0.25, F-Score 0.375, **AER 0.5**

- AER überschätzt systematisch die Rolle von Precision
  - Konzentration auf "sure links" führt zu hohem 1-AER
    - Alignments mit hohem 1-AER **sind oft sehr spärlich**

Wird es besser mit F-Score?



Generalisierung von F-Score

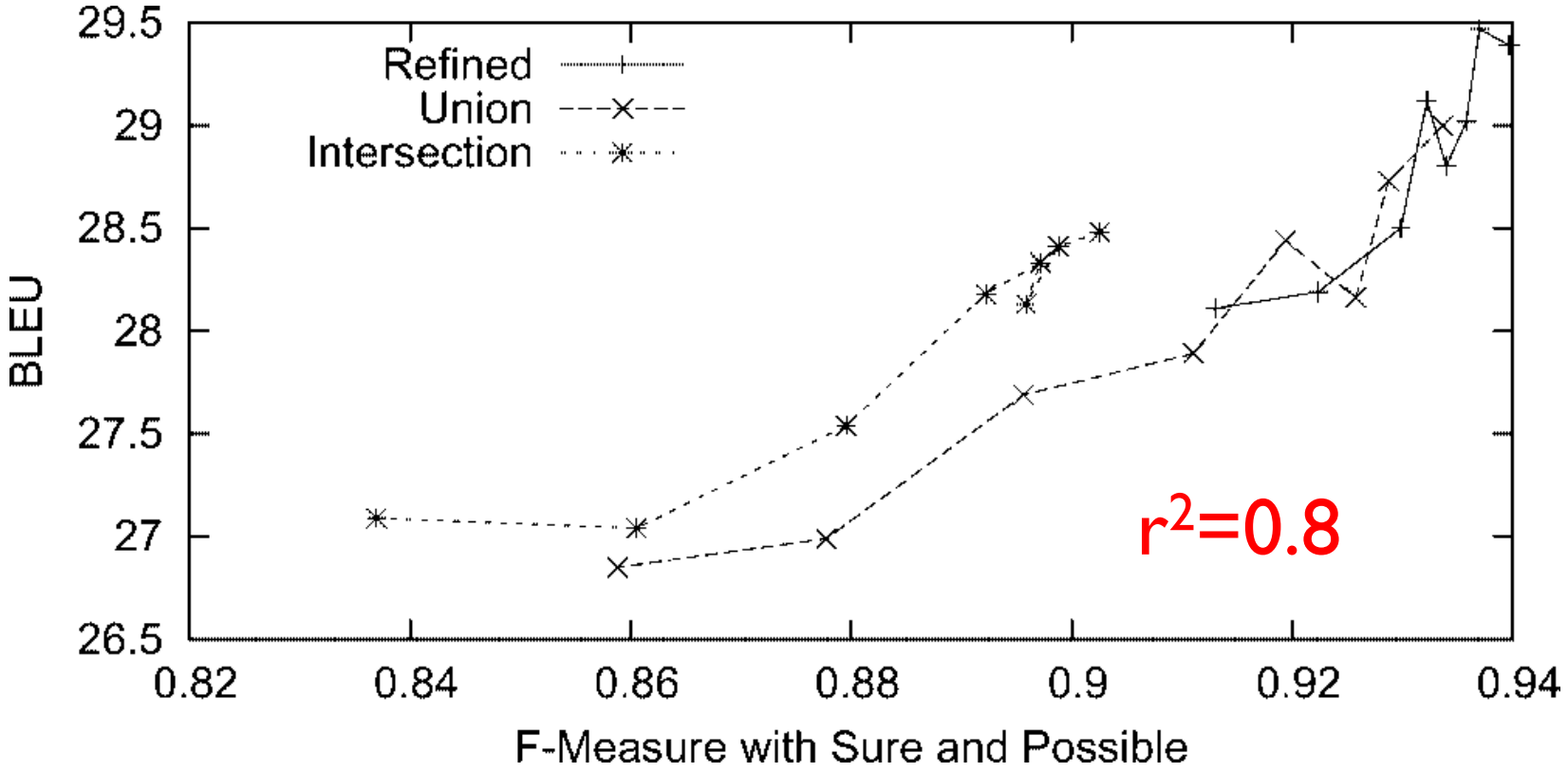
- Wir brauchen eine allgemeinere Erklärung für den Zusammenhang zwischen Alinierungsqualität und Übersetzungsqualität
  - Übersetzungsqualität hängt vom Verhältnis Precision/Recall ab
  - Aber perfekte Balance ist auch nicht ideal

- α-F-Score: Erweiterung des F-Scores um einen Parameter, der das "ideale" Verhältnis Precision/Recall festlegt

$$\alpha\text{-F} = \frac{Pr * R}{(1-\alpha)Pr + \alpha R}$$

- α liegt zwischen 0 und 1
- α=0.5: Gleichgewicht
- α=0.1: Recall 9x so wichtig wie Precision

F-Score mit α=0.1



Lehren

1. Nicht AER verwenden!
   - AER verhält sich unsystematisch

2. Für MÜ ist Recall scheinbar deutlich wichtiger als Precision
   - Spärlichkeit ist ein größeres Problem als Rauschen