Beispiel-basierte Maschinelle Übersetzung

Cristina Vertan

Terminologie und Abkürzungen

- Ein Korpus ist eine Sammlung von Texten in einer Domäne, die die wesentlichen Sprachmerkmale einer gegebenen Domäne enthält. Die übliche Größe ist ab 10.000 Wörter.
- Ein paralleles Korpus ist eine Sammlung von Texten und ihren Übersetzungen in 2 oder mehr Sprachen.
- Eine Korpusannotation ist die Markierung von festgelegten Sprach- oder Ausdrucksmerkmalen in einem Korpus.
- EBMT = Example based Machine Translation
- TM = Translation Memory

Inhalt

- Prinzipien und die Architektur von EBMT-Systemen
- Wie baut man eine Übersetzungsbeispiel-Datenbank?
- Analyse der Eingabe
- Datenbanksuche
- Komposition der Ausgabe
- EBMT und andere MT-Systemtypen
- Was implementieren wir im Praktikum?

„A good translator is a lazy translator“

Quellen der EBMT: Übersetzungstheorie

Eine neue Übersetzung muss soviel Material wie möglich aus bisherigen Übersetzungen (derselben Domäne, Zeit, etc) benutzen.

Vorteile: Dieser Mechanismus

- spart Zeit

- sichert die terminologische und stilistische Konsistenz

Viele Übersetzungen entstehen auch bei Humanübersetzern einfach durch Änderungen in bereits existierenden Übersetzungen

Analogy principle (Nagao, 1984)

Quellen der EBMT: Kognitionswissenschaft

- Die menschlichen Übersetzungen von einfacheren Sätze sind nicht das Ergebnis einer tiefen linguistischen Analyse, sondern eher einer passenden
  - Zerlegung des Satzes in Teilkomponenten, gefolgt von
  - Übersetzung der Teilkomponenten, sowie der
  - Kombination dieser Übersetzungen.
- Die Übersetzung der Teilkomponenten wird durch Analogie mit anderen schon bekannte Übersetzungen gemacht.

Quellen der EBMT: MAHT

- Übersetzer benutzen oft große Datenbanken mit Übersetzungsbeispielen (Translator‘s workbenches /Translation memories).
- Z.B. TRADOS - ein TM-System für 12 europäische Sprachen.
- Das System sucht in der Datenbank Einträge die mit der Eingabe ähnlich sind und zeigt ihre Übersetzungen.
- Der Übersetzer macht ausschließlich die Auswahl von Teilen, die er braucht, sowie deren Rekombination

Funktionalität eines EBMT-System

- Man extrahiert aus einem parallelen Korpus „relevante“ Beispiele und speichert sie in einer Datenbank.
- Die Eingabe wird mit den Einträgen in der Datenbank verglichen (matching-Phase).
  - Entweder versucht man, (Teile der) Eingabe in der Datenbank identisch zu finden oder
  - Bei Nichtidentität berechnet man einen Abstand zwischen Datenbankeinträgen und (Teilen der) Eingabe und benutzt die Einträge mit dem kleinsten Abstand.
- Dann muß jeweils der übereinstimmende Teil des zielsprachlichen Ausdrucks extrahiert werden (alignment-Phase); das ist trivial für Identität
- Die entsprechenden Teilübersetzungen aus der Datenbank werden zu einer korrekten Übersetzung rekombiniert.

Relevante Beispiele?

- Geeignet für eine gute lexikalische Abdeckung: Viel domänenrelevanter Wortschatz
  - Möglichst mit Kookkurrenzen (reflexiv, Partikelverben)
- Geeignet für eine gute syntaktische Abdeckung, z.B. für dt.
  - Strukturen in Haupt- und Nebensatzfolge
  - Aktiv- und Passivsätze
  - Frage- und Aussagesätze
  - Mit verschiedenen eingebetteten Strukturen, z.B. Attributsatz, Inhaltssätze, Konjunktionalsätze

EBMT - Beispiel

- Eingabe: *Ungeeigneter Kraftstoff kann zu Motorschäden führen*

- Die Übersetzungsbeispiele-Datenbank enthält:

- Starke Motorbelastung kann zu Motorschäden führen - High engine loading can cause engine damage

-Ungeeigneter Kraftstoff darf nicht benutzt werden.- Unsuitable fuel must not be used

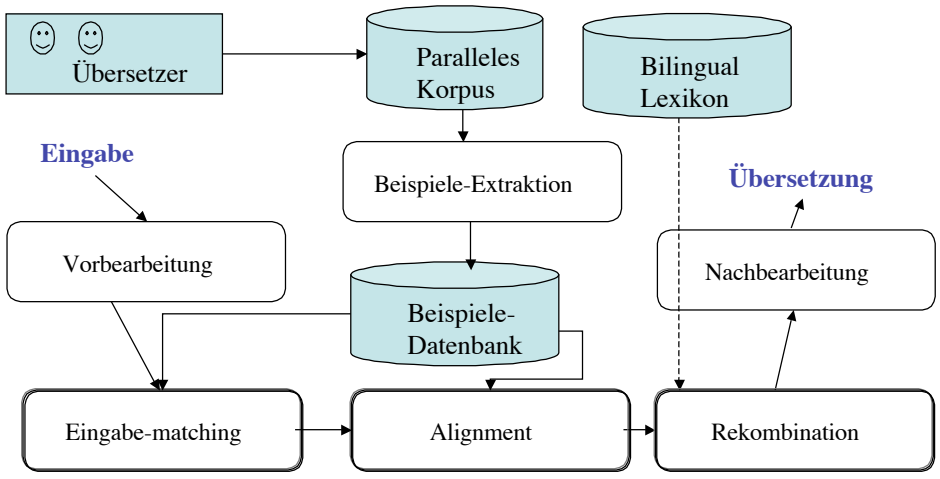
- Man identifiziert folgende Segmente.

- kann zu Motorschäden führen - can cause engine damage.

-Ungeeigneter Kraftstoff - Unsuitable fuel

- Die Übersetzung ist dann:
  - *Unsuitable fuel can cause engine damage*

Architektur der EBMT-Systeme



Inhalt

- Prinzipien und die Architektur von EBMT-Systemen
- Wie baut man die Übersetzungsbeispiel-Datenbank
- Analyse der Eingabe
- Datenbanksuche
- Komposition der Ausgabe
- EBMT und andere MT-Systemtypen
- Was implementieren wir im Praktikum?

Wie baut man eine Übersetzungsbeispiele-Datenbank

Wichtige Entscheidungen:

- Umfang: Wie viele Übersetzungsbeispiele sollen gespeichert werden?
- Länge der Einträge: Wie lang sollen die Beispiele sein ?
- Annotationen: Braucht man zusätzliche Information?
- Daten: Was speichert man (Strings, grammatische Strukturen)?

Größe der Datenbank und Beispiel-Länge

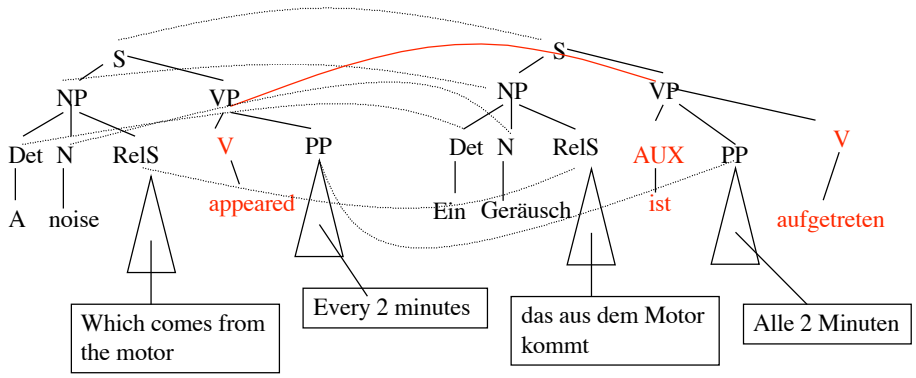
- Die Größe der Übersetzungsbeispiele-Datenbank variiert zwischen ein paar hundert und 800 000 Einträge.
- Je größer die Datenbank ist, desto besser funktioniert das System.
- Es gibt keine ideale Länge, die die Beispiele haben sollten:
  - Je länger die Beispiele sind, desto niedriger ist die matching-Quote
  - Je kürzer die Beispiele sind, umso größer ist die Chance, Ambiguitäten (mehrere mögliche Übersetzungen) zu erzeugen
- Die Standard-Beispielslänge ist der Satz.

Zusätzliche Annotation der Beispiele

- Die Einträge in der Datenbank haben die Form:
  $Q_1,Q_2,....Q_i...,Q_n$ □ $Z_1,Z_2,....Z_j,...Z_m$ wobei Q_i und Z_i = Wörter
- Die Schwierigkeit in der Alignmentphase ist fest zu stellen, ob z.B. der Teilausdruck Z_1Z_2 die Übersetzung von des Teilausdrucks Q_1,Q_2 ist, ohne dass man rekursiv einen Zerlegungs-/Konstruktions-mechanismus aufruft ().
- Um diese Operation zu unterstützen, kann man Morpheme, die einen klaren Kontext markieren, entsprechend annotieren:
- Beispiel solcher Morpheme: Quantoren, Konjunktionen, Fragepartikel, usw.
- Z.B. <QUANT> all uses ... </QUANT>
  <QUANT> alle Benutzungen ... </QUANT>

Datenbank mit grammatischen Strukturen

Die Übersetzungsbeispiele sind nicht mehr Strings sondern syntaktische Patterns mit entsprechenden links.



Inhalt

- Prinzipien und Architektur von EBMT-Systemen
- Wie baut man die Übersetzungsbeispiel-Datenbank?
- Matching der Eingabe
- Datenbanksuche
- Komposition der Ausgabe
- EBMT und andere MT-Systemtypen
- Was implementieren wir im Praktikum?

Matching - Allgemeines

- Man muss in der Datenbank die Einträge suchen, die am ähnlichsten mit der Eingabe sind.
- Es gibt zwei Probleme:
  - Wie misst man die Ähnlichkeit?
  - Wie durchsucht man die Datenbank?
- Für die Berechnung der Ähnlichkeit benutzt man normalerweise eine Kombination von stringbasierten und statistisch basierten Methoden.

Zeichenbasiertes Matching

- Die Ähnlichkeit wird aufgrund der Zeichen in der Eingabezeichenkette (wortweise) berechnet. Folgende Abstände werden benutzt:
  - "Längste gemeinsame Teilkette"
  - "Edit distance": wieviele Operationen (Einfügen, Löschen, Ersetzen) sind nötig, um die Eingabekette zu der Kette aus der Datenbank zu machen
- Diese Methoden lassen sich einfach durch Greedy oder Dynamische Programmierungs-Verfahren implementieren

Zeichenbasiertes matching - Probleme

- Die inhaltliche Ähnlichkeit wird nicht erkannt.
- Z.B. : Nach diesen Kriterien ist
  - *(1) Haben sie einen Rat?* näher an
  - *(2) Haben Sie ein Rad?* als
  - *(3) Haben sie einen Hinweis?*

- Die "Edit-distance" zwischen (1) und (2) ist 3, während die "edit-distance" zwischen (1) und (3) 7 ist!

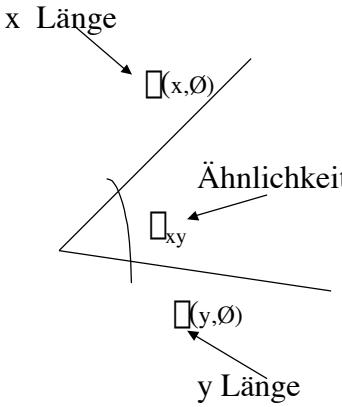
Wortbasiertes matching -1-

- Man benutzt zusätzlich einen sog."Thesaurus", eine Art semantisches Netz von Wörtern in dem man deren inhaltliche Nähe in Bezug auf Bedeutung und Verwendung notiert.
- Z.B. gegeben im Beispiel Datenbank:
  - *Der Abstand zwischen den Kontrollen soll 2 Jahre nicht überschreiten → The interval between 2 general checks should not exceed 2 years.*
  - *Der normale Abstand zwischen den Nebelleuchten ist x cm. → The normal distance between fog-lights is x cm.*
- Dann die Eingabe: *Wo finde ich den Abstand zwischen den Rädern?*
  - *Räder* ist im Thesaurus näher zu *Nebelleuchten,* deswegen wird *Abstand* mit *Distance* übersetzt.
- Obwohl die edit distance zwischen *Räder* und *Kontrolle* kleiner ist als die edit distance zwischen *Räder* und *Nebelleuchte*.

Wortbasierte Matching -"Angle of similarity"-1

- Man benutzt eine trigonometrische Distanzmessung
- Der Abstand zwischen zwei Sätzen wird durch eine Differerenzfunktion berechnet.
- Diese Differenzfunktion funktioniert ähnlich wie das zeichenbasierte matching (man zählt die Anzahl der Operationen).
- Die Operationen sind jedoch gewichtet, z.B. das Hinzufügen eines Kommas hat weniger Gewicht als ein fehlendes Adjektiv.
- Die Gewichte sind nicht festgelegt, sondern werden abhängig von System und Übersetzungsdomäne gewählt.

Wortbasierte Matching -"Angle of similarity"-2



Abstand zwischen Satz x und Satz y

$$\sin\frac{\alpha_{xy}}{2} = \frac{\alpha(x,y) - |\alpha(x,\varnothing) - \alpha(y,\varnothing)|}{2 \cdot \min\{\alpha(x,\varnothing),\alpha(y,\varnothing)\}}$$

Wortbasiertes Matching - ”Angle of similarity” Beispiel

1. *Lesen Sie Seite 3 im Kapitel “Benzin”*
2. *Lesen Sie Seite 3 im Kapitel “Benzin” und Seite 5 in Kapitel “Länderspezifische Bemerkungen”*
3. *Lesen Sie Seite 4 im Kapitel “Bremsen”.*

- Zeichenbasiertes matching ergibt eine größere Ähnlichkeit von Satz 1 mit Satz 3 weil nur 1 Wort unterschiedlich ist.
- Aber: Satz 2 ist eigentlich eine bessere Wahl, da dort Satz 1 vollständig enthalten ist. Diese Wahl wird von der “angle distance” errechnet.

Matching - andere Verfahren

- Annotiertes wort-basiertes matching: Man berücksichtigt auch Annotationen im Satz und entsprechende Markierungen im Lexikon
- Partielles matching: Man vergleicht die Eingabe grundsätzlich nicht mit ganzen Beispielen in der Datenbank sondern nur mit Teilen davon
- Stuktur-basiertes matching: Vergleich von Baumstrukturen

Inhalt

- Prinzipien und Architektur von EBMT-Systemen
- Wie baut man die Übersetzungsbeispiele-Datenbank
- Analyse der Eingabe
- Datenbanksuche
- Komposition der Ausgabe
- EBMT und andere MT-Systemtypen
- Was implementieren wir im Praktikum?

Datenbank-Suche (Alignment) -1-

- Im Idealfall findet man in der Datenbank ein Beispiel, das mit der Eingabe identisch ist.
- Normalerweise muss man aus den Übersetzungsbeispielen passenden Fragmente herausschneiden
- Je einfacher die Datenbank strukturiert ist (kein PoS - alignment, keine zusätzlichen Markierungen), desto schwieriger ist dieser Schritt.

Datenbank-Suche (Alignment) -2-

- Es gibt statistische Verfahren die das alignment automatisch machen. Diese Verfahren basieren auf komplizierten statistischen Modellen der Quell- und Zielsprache.
- Einfacher: man speichert die syntaktische Struktur der Beispiele in Quell- und Zielsprache sowie die entsprechenden Verbindungen.
- Man identifiziert dann Teilstrukturen.

Inhalt

- Prinzipien und Architektur von EBMT-Systemen
- Wie baut man die Übersetzungsbeispiele-Datenbank
- Analyse der Eingabe
- Datenbanksuche
- Komposition der Ausgabe 
- EBMT und andere MT-Systemtypen
- Was implementieren wir im Praktikum?

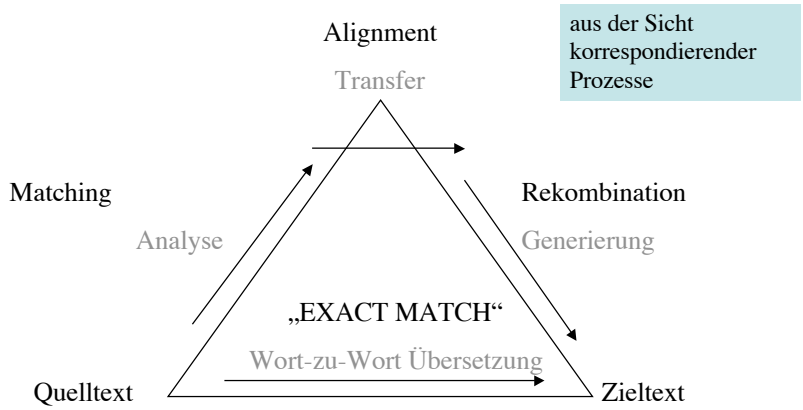
Komposition der Ausgabe

- Wenn man ohne grammatische Stukturen arbeitet, ist das der schwierigste Lösungsweg,
- Wenn man Teilbäume hat, ist der Prozess eigentlich nur eine Unifikation von Bäumen
- Für stark flektierende Sprachen muss man die Wahl der Artikelform z.B. durch statistische Berechnungen in großen Korpora feststellen, oder ist es Teil der Nachbearbeitung.

Inhalt

- Prinzipien und Architektur von EBMT-Systemen
- Wie baut man die Übersetzungsbeispiele-Datenbank
- Matching der Eingabe
- Datenbanksuche
- Komposition der Ausgabe
- EBMT und andere MT-Systemtypen 
- Was implementieren wir im Praktikum?

Übersetzungsdreieck für regelbasierte Verfahren vs. EBMT



@A. Way

EBMT and Translation Memories

- Die Analyse- und Durchsuchungsmechanismen sind identisch
- Bei TM gibt es keine Rekombinationsphase.
- Translation Memories sind nur als Unterstützungs-werkzeuge für Übersetzer gedacht. Deswegen werden die Übersetzungen, die in der Datenbank gefunden werden, dem Übersetzer nur präsentiert und er trifft die richtige Auswahl.

EBMT und statistische MT

- Es gibt Versuche, die zwei Verfahren zu kombinieren.
- Ausnahmen, die nicht statistisch modelliert werden können oder sehr häufige Ausdrücke werden durch EBMT-Verfahren übersetzt.
- Z.B. im Verbmobil System:
  - Ich denke mal - I think
  - Also wir wollten - well, we plan to
  - In unserer Abteilung ein neues Netzwerk aufbauen - set up a new network in our department.

Vergleich von linguistischen und empirischen Methoden

- Im Verbmobil-System (Deutsch-Englisch-Japanisches Speech-to-Speech Translation System) wurden 4 MT Verfahren implementiert :
  - Ein transferbasiertes,
  - ein statistisches und
  - ein beispielbasiertes
  - [ein dialogbasiertes].
- Nach der Evaluation war die Anzahl von Sätzen, die inkorrekt übersetzt worden waren, bei
  - Semantischem Transfer 62 %
  - Beispielbasierter MT 35%
  - Statistischer MT 29%

Aktuelle Trends in EBMT

- Automatische Extraktion von Übersetzungspatterns :
  - aus einem parallelen Korpus,
  - aus Dependenz-Strukturen.
- Integration von EBMT und statistischer maschineller Übersetzung
- Integration von linguistischen Verfahren in EBMT.

Inhalt

- Prinzipien und Architektur von EBMT-Systemen
- Wie baut man die Übersetzungsbeispiele-Datenbank
- Analyse der Eingabe
- Datenbanksuche
- Komposition der Ausgabe
- EBMT und andere MT-Systemtypen
- Aktuelle Trends in EBMT
- Was implementieren wir im Praktikum? 

Was implementieren wir im Praktikum?

- Die Beispiele in der Übersetzungsdatenbank werden zusammen mit den entsprechenden Baumstrukturen gespeichert.
- Wir werden verschiedene matching-Verfahren implementieren und evaluieren:
  - Zeichenbasierte Verfahren
  - Wortbasierte Verfahren.
- Das Alignment und die Rekombinationschritte werden sich auf Baumoperationen stützen.
- Keine morphologische Nachbearbeitung