Example Based Machine Translation

Beispielbasierte maschinelle Übersetzung

Einführung (1)

Die Beispiel-basierte ist eine Möglichkeit der maschinellen Übersetzung.

Der Grundgedanke besteht darin, neue Übersetzungen mit Hilfe einer Datenbank von (guten) Übersetzungen zu generieren.

Während die meisten EBMT Systeme ab ca. 1990 entstanden, hat der beispiel-basierte Ansatz einige Vorläufer. Die wichtigsten Ideen, die EBMT beeinflusst haben, sind „Translation Memory“ und „Analogie-basierte Übersetzung“.

Einführung (2)

Translation Memory (TM) schlägt einen Editor für Übersetzer vor, der über eine Datenbank von früheren Übersetzungen verfügt und diese für den Übersetzer, der mit dem Editor arbeitet, nach „Präzedenzfällen“ untersucht.

Somit wird der Übersetzer unterstützt (TM generiert aber nicht selbst eine Übersetzung). Die Idee von TM wird Martin Kay zugeschrieben, er hat sie in seinem Paper „Proper Place“ 1980 formuliert.

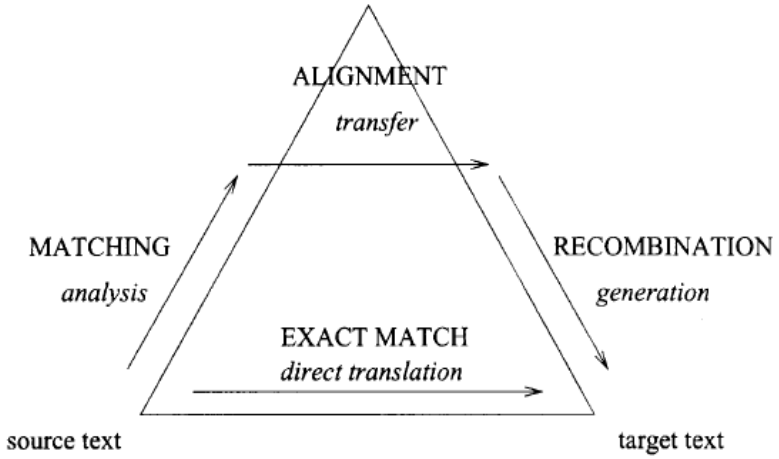
Die Analogie-basierte Übersetzung geht auf Makoto Nagao zurück. Er beschreibt in einem Paper von 1984 jene Vorgänge als Analogie-basiert, die heute auch für EBMT gelten.

Modell (1)

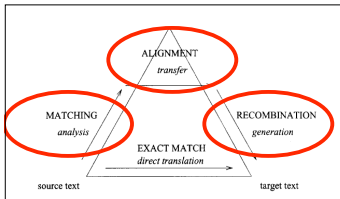
•MAKOTO NAGAO (1984):

> «Man does not translate a simple sentence by doing deep linguistic analysis, rather, [...] first, by properly decomposing an input sentence into certain fragmental phrases ..., then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as ist reference.»

Modell (2)



Blockseminar Computerlinguistik
Universität Basel, SS 2004 5



Modell (3)

Matching

Suche nach passenden Beispielen für den Eingangstext, «Suchproblem».

‚to match' = passen zu, Passendes finden

Alignment

Aus den gefundenen Beispielübersetzungen die passenden Fragmente auswählen.

‚alignment' = Ausrichtung

Recombination

Die aus der Übersetzung ausgewählten Fragmente zum Zieltext zusammenfügen.

Blockseminar Computerlinguistik
Universität Basel, SS 2004 6

Datenbasis (1)

Woher eine Beispieldatenbank nehmen, die zwei- oder mehrsprachig ist?

- Der Korpus wird von Menschen erarbeitet.
  - –Sehr zeitaufwändig und teuer
  - –Dadurch ist die mögliche Grösse dieser Datenbanken ziemlich beschränkt, und damit auch ihr Anwendungsgebiet.

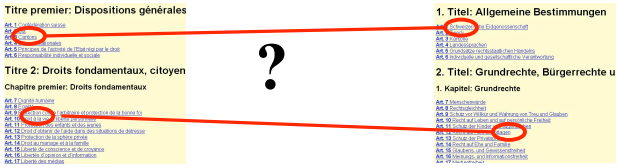
Blockseminar Computerlinguistik
Universität Basel, SS 2004 7

Datenbasis (2)

Die andere Möglichkeit besteht darin, einen schon bestehenden Korpus als Datenbasis zu nehmen:

- –Mehrsprachige Internetseiten
- –Gesetzes- und andere offizielle Texte von mehrsprachigen Ländern und Organisationen

Was in einem solchen Korpus jedoch fehlt, sind Informationen über die Korrespondenz der einzelnen Teile (Sätze, Redeweisen, Wörter, etc.).



Blockseminar Computerlinguistik
Universität Basel, SS 2004 8

Datenbasis (3)

Bevor der Korpus eingesetzt werden kann, müssen diese Informationen hinzugefügt werden, der Korpus muss „aligned“ werden. Es gibt teure („resource-rich“) und günstige („resource-poor“) automatische Alignment-Verfahren.

Günstig:

- Statistik über die Satzlängen
- Statistik über die Häufigkeit identischer Wörter innerhalb des Satzes
- (Eingeschränkte) lexikalische Information

Teuer:

- Einsatz von zweisprachigen Lexika und Glossaren – d.h. es wird versucht, inhaltliche (und nicht rein statistische) Korrespondenzen zu entdecken.

Datenbasis (4)

Je nach Granularität der Beispiele ist das Alignment unterschiedlich schwierig:

- –Korrespondierende Absätze sind aus der Textstruktur relativ klar ersichtlich
- –Korrespondierende Sätze sind durch günstige Methoden (meist) identifizierbar.
- –Korrespondierende Wörter können mithilfe teurer Verfahren identifiziert werden.
- –Problematisch sind Redewendungen, Mehr-Wort-Begriffe, Kollokationen.

Datenbasis (5)

Es scheint, dass Übersetzungsbeispiele, die kürzer als ein Satz, aber länger als ein Wort sind, schwieriger zu erhalten sind.

Die Länge der Beispiele hat in zweierlei Hinsicht Einfluss auf die Übersetzung:

> Je grösser die Beispiele, desto geringer die Ambiguität – desto geringer aber auch die Anzahl an Eingaben, die abgedeckt sind.
>
> Je kleiner die Beispiele, desto grösser die Ambiguität – desto grösser aber auch die Anzahl an Eingaben, die abgedeckt sind.

Datenbasis (6)

Das Alignment ist uns hier im Zusammenhang mit der Erstellung des Korpus begegnet.

Wie wir in der (EB)MT-Pyramide gesehen haben, spielt es als mittlerer Schritt der EBMT eine weitere Rolle – wir werden später darauf kommen.

Datenbasis (7)

In Bezug auf die Beispieldatenbank kommen wir jetzt zur Speicherungsform.

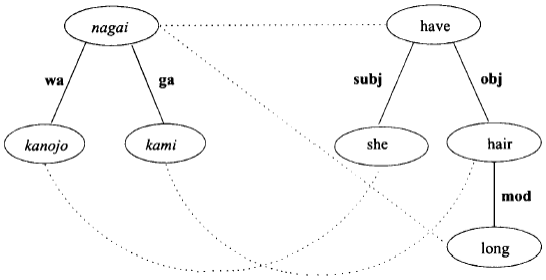
Am einfachsten ist es, Beispiele in Form von korrespondierenden Zeichenketten zu speichern.

«Ich denke, also bin ich» = «Cogito ergo sum»

Da keine zusätzliche Information gespeichert wird, könnte das System jedoch den Satz «Ich zweifle, also bin ich» nicht ohne weitere Analyse anhand dieses Beispiels übersetzen.

Datenbasis (8)

In frühen EBMT Systemen wurden die Beispiele als „annotated trees“ – kommentierte Bäume – gespeichert. Liegt der Eingangstext auch in dieser Form vor, können Unterschiede spezifisch erkannt und behandelt werden.



Datenbasis (9)

Eine weitere Möglichkeit sind „Verallgemeinerte Beispiele“. Sie haben die grundsätzliche Struktur von wörtlichen Beispielen (korrespondierende Zeichenketten), beinhalten aber neben normalen Wörtern auch Variablen:

„(X) ergo sum“ = „(X), also bin ich“

Oft sind die Variablen dabei typisiert, d.h. verraten etwas über die Art des vertretenen Satzteils (Nummer, Name, Datum, Wortart).

Datenbasis (10)

Neben Beschaffung der mehrsprachigen Texte, Alignment und Speicherungsform gibt es weitere Aspekte in Bezug auf die Datenbasis von EBMT. Bevor wir auf den in der (EB)MT Pyramide gezeigten Übersetzungsvorgang eingehen, möchte ich davon einen noch erwähnen:

Der Korpus, der einem EBMT System zur Verfügung steht, *definiert implizit dessen Subsprache*. Die Wahl der Übersetzungsbeispiele wirkt sich auf das ideale Anwendungsfeld aus.

Matching (1)

•Die Suche nach einem geeigneten Beispiel ist eine Distanz- respektive Ähnlichkeitsmessung.

•Je grösser die Distanz zwischen zwei Texten, desto geringer die Ähnlichkeit.

•Es gibt verschiedene Kriterien für diese Messung, die jeweils mehr oder weniger „linguistisch“ sind.

Matching (2)

Zeichenbasiert (character-based)

Die Ähnlichkeit wird aufgrund der Zeichen in den Zeichenketten bestimmt. Mögliche Kriterien sind dabei:

„Längste gemeinsame Teilkette“: „Discordia“ ist näher an „Concordia“ als „Concerdia“ („cordia“ länger als „conc“ und „rdia“)

„Edit distance“: wieviele Operationen aus (Einfügen,Entfernen,Ersetzen) sind nötig, um Identität herzustellen? Beispiel: „sehn“ ist von „sehen“ gleich weit entfernt wie „stehen“, nämlich 1 Einfügen respektive 1 Entfernen.

Matching (3)

Zeichenbasiert (character-based)

Problem: inhaltliche Ähnlichkeit wird nicht erkannt.

(a)«Es spielt eine Rolle» ist näher an (b) «Er spielt Roulette» als an (c) «Es ist bedeutungsvoll»

Teilkette: „spielt“ länger als „Es“
Distanz: (a) → (b) = 9 Operationen, (a) → (c) = 24 Operationen

Matching (4)

Wortbasiert (word-based)

Die Distanz wird festgestellt, indem eine Datenbank konsultiert wird, die die inhaltliche Nähe von Wörtern in Bezug auf Bedeutung und Verwendung enthält.

Wir haben folgende zwei Übersetzungsbeispiele:

A man eats vegetables → Ein Mann isst Gemüse.
Acid eats metal → Säure zerfrisst Metall.

Und folgenden zu übersetzenden Text:

He eats potatoes.

Die Datenbank gibt Auskunft darüber, dass „He“ näher an „A man“ als an „Acid“ ist, und dass „potatoes“ näher an „vegetables“ als an „metal“ ist, also wählt das System die Übersetzung:

Er isst Kartoffeln. (Und nicht: Er zerfrisst Kartoffeln)

Matching (5)

„Angle of similarity“ (Carroll)

Diese Distanzmessung ist eine trigonometrische.

Es existiert eine Differenzfunktion □, die den Abstand zwischen zwei Sätzen angibt. Diese Differenzfunktion funktioniert ähnlich wie die beim zeichenbasierten Matching vorgestellten (Anzahl Operationen), gewichtet die einzelnen Operationen jedoch: das Hinzufügen eines Kommas ist z.B. weniger gewichtig als ein fehlendes Adjektiv.

Um zwei Sätze <x,y> zu vergleichen, werden drei Werte ermittelt: □(x,y), □(x, ø) und □(y, ø), wobei ø der Nullsatz („“) ist. Diese drei Werte sind die Seitenlängen eines Dreiecks.

Blockseminar Computerlinguistik
Universität Basel, SS 2004 21

Matching (6)

„Angle of similarity“ (Carroll)

Der Winkel zwischen <x,y> (zwischen den Seiten □(x,ø) und □(y,ø)) gibt nun Aufschluss über die „wahre Differenz“ von <x,y>: ist er sehr klein (oder Null), ist die Differenz sehr klein.

Das Verfahren ist relativ kompliziert, erreicht jedoch, dass unterschiedliche Satzlängen weniger ins Gewicht fallen als andere Differenzen.

So hätte „Es geht mir gut“ zu „Es geht mir schlecht“ eine grössere Distanz als zum längeren Satz „Es geht mir gut, danke für die Nachfrage“.

Blockseminar Computerlinguistik
Universität Basel, SS 2004 22

Matching (7)

Es gibt noch weitere Matching-Techniken, wie „Annotated Word-based-matchin“, „Stucture-based matching“, „Partial matching for coverage“, die ich hier aber nicht detailliert vorstellen möchte.

Alignment (1)

Im Idealfall finden sich Beispiele, deren „Eingangstext“ mit dem zu Übersetzenden Eingangstext identisch ist. Ob dies wahrscheinlich ist, hängt natürlich von der vorhin angesprochenen Granularität der Beispiele ab.

Tritt dieser Fall nicht ein, müssen aus den Übersetzungsbeispielen passende Teilstücke (Fragmente) ausgewählt werden.

Wir sehen: wurde beim Aufbau des Korpus „gespart“, sprich sind die Beispiele relativ gross und in trivialer Form gespeichert, muss hier einige Arbeit nachgeholt werden.

Alignment (2)

Wir möchten den folgenden Satz übersetzen:

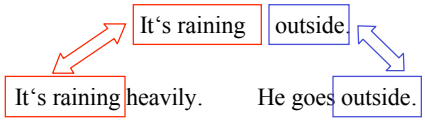
It‘s raining outside.

Die Beispiele, die wir im *Matching* ausgesucht haben, sind:

He goes outside → Er geht nach draussen.

It‘s raining heavily → Es regnet stark.

Alignment bedeutet nun, dass die sich deckenden Teile in Eingangstext und Beispielen identifiziert werden müssen:



Blockseminar Computerlinguistik
Universität Basel, SS 2004 25

Alignment (3)

Nachdem die Beispielteile identifiziert sind, muss noch das ihnen korrespondierende Fragment in der Übersetzung identifiziert werden:

It‘s raining heavily → Es regnet stark.

He goes outside → Er geht nach draussen.

Blockseminar Computerlinguistik
Universität Basel, SS 2004 26

Recombination (1)

Nachdem Eingangstext und Beispiele derart „ausgerichtet“ sind, ist die Generation des Zieltextes der letzte Schritt in der Übersetzung.

Es ist (meist) nicht möglich, die „ausgerichteten“ Fragmente aus dem *Alignment* einfach aneinander zu hängen; wir erhielten im eben dargestellten Fall dann nämlich:

It‘s raining outside → *(Es regnet nach draussen)

Blockseminar Computerlinguistik
Universität Basel, SS 2004 27

Recombination (2)

Die Textgeneration, die *Recombination*, muss also sicherstellen, dass die Fragmente der Zielsprache nicht irgendwie, sondern konform zur Grammatik zusammengesetzt werden:

It‘s raining outside → Es regnet draussen.

Blockseminar Computerlinguistik
Universität Basel, SS 2004 28

Recombination (3)

In ihrem Aufsatz «[EBMT]: A New Paradigm» meinen Chunyu Kit, Haihua Pan und Jonathan Webster:

> «[Recombination] is recognized as the most difficult step in EBMT»

Sie gehen davon aus, dass die Erzeugung eines gut lesbaren, grammatikalischen Zieltextes nicht durch Grammatikregeln gewährleistet werden kann – dies widerspreche dem empirischen Ansatz, den EBMT verfolgt. Sie wählen ein statistisches Modell.

Recombination (4)

Für die Zielsprache existiert ein statistisches Modell. Für eine bestimmte Wortfolge <x,y,z> gibt es Auskunft über deren Häufigkeit – ergibt sich aus dem Zusammensetzen der Fragmente eine höchst seltene oder nie vorkommende Wortfolge, ist es höchstwahrscheinlich, dass sie schlecht lesbar oder ungrammatikalisch ist.

In seinem «Review Article:[EBMT]» beschreibt Harold Somers ein konkretes Problem, die „boundary friction“, die er durch eine Suchabfrage und die Berücksichtigung der gefundenen Treffer löst.

Recombination (5)

Im Deutschen (und vielen anderen Sprachen) wirken sich grammatikalische Fälle auf die Form von z.B. Artikeln aus. Dies ist z.B. im Englischen nicht der Fall. Die daraus resultierenden Schwierigkeiten werden auch als „boundary friction“ (~ Grenzreibung) beschrieben.

Das Englische „the“ kann im Deutschen mit „der“, „den“ o.a. übersetzt werden.

I saw the handsome boy

* Ich sah der schöne Junge.
Ich sah den schönen Jungen.

Recombination (6)

Das System muss im statistischen Ansatz nun den Fall, der für die Wahl der Artikelform relevant ist, nicht kennen: es misst schlicht die Häufigkeit der beiden (oder mehreren) Übersetzungsmöglichkeiten.

Somers demonstriert dies durch eine Suchabfrage bei eine Suchmaschine. Bei Google ergab dies:

„Ich sah den“ ergab 8010 Treffer
„Ich sah der“ ergab 454 Treffer (es handelt sich dabei vorwiegend um poetische Genitive wie „Ich sah der Liebe Licht“)

Aufgrund der Trefferzahl erscheint klar, welche Übersetzung das System zu wählen hat.

Einsatz von EBMT (1)

EBMT wird selten als stand-alone System eingesetzt.

Es bildet meist den Teil eines Übersetzungssystems, das sich mehrerer Techniken bedient.

Für EBMT sprechen insbesondere:

- –Wiederverwendbarkeit der Übersetzungen
- –Es werden menschliche Übersetzungen verwendet, der Anteil an „Computersprache“ in einer Übersetzung ist geringer als bei anderen MT
- –Einfache Erweiterbarkeit (Beispiele hinzufügen ←→ Regeln hinzufügen)
- –Geringere Gefahr der Strukturübertragung von Sprache 1 auf Sprache 2
- –Relativ geringere Kosten zur Erstellung eines neuen Systems

Einsatz von EBMT (2)

System	Reference(s)	Language pair	Size
PanLite	Frederking & Brown (1996)	Eng → Spa	726 406
PanEBMT	Brown (1997)	Spa → Eng	685 000
TDMT	Sumita et al. (1994)	Jap → Eng	100 000
CTM	Sato (1992)	Eng → Jap	67 619
Candide	Brown et al. (1990)	Eng → Fre	40 000
no name	Murata et al. (1999)	Jap → Eng	36 617
PanLite	Frederking & Brown (1996)	Eng → SCr	34 000
TDMT	Oi et al. (1994)	Jap → Eng	12 500
TDMT	Mima et al. (1998)	Jap → Eng	10 000
no name	Matsumoto & Kitamura (1997)	Jap → Eng	9 804
TSMT	Sobashima et al. (1994)	Eng → Jap	607
TDMT	Furuse & Iida (1992a, b, 1994)	Jap → Eng	500
TTL	Öz & Cicekli (1998)	Eng ↔ Tur	488
TDMT	Furuse & Iida (1994)	Eng → Jap	350
EDGAR	Carl & Hansen (1999)	Ger → Eng	303
ReVerb	Collins et al. (1996), Collins & Cunningham (1997), Collins (1998)	Eng → Ger	214
ReVerb	Collins (1998)	Irish → Eng	120
METLA-1	Juola (1994, 1997)	Eng → Fre	29
METLA-1	Juola (1994, 1997)	Eng → Urdu	7

Key to languages – Eng: English, Fre: French, Ger: German, Jap: Japanese, SCr: Serbo-Croatian, Spa: Spanish, Tur: Turkish

Bibliographie

Somers, Harold: **Review Article: Example-based Machine Translation**. In: Machine Translation 14: 113-157. Dordrecht: Kluwer Academic Publishers, 1999.
(verfügbar unter http://stp.ling.uu.se/~ebbag/somers.pdf)

Kit,Pan & Webster: **Example-Based Machine Translation: A New Paradigm**. In S.W. Chan (ed.), *Translation and Information Technology,* pp.57-78. Hong Kong: Chinese U of HK Press, 2002.
(verfügbar unter http://personal.cityu.edu.hk/~ctckit/papers/EBMT-review-CUHK.pdf)

Diverse Aufsätze im Zuge des **8th Machine Translation Summit**, verfügbar unter: http://iai.uni-sb.de/~carl/ebmt-workshop/papers.html