

Softwareprojektpraktikum Maschinelle Übersetzung

Matthias Huck, Markus Freitag {huck,freitag}@i6.informatik.rwth-aachen.de

Vorbesprechung 2. Aufgabe 28. April 2011

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany





Was ist statistische maschinelle Übersetzung?

"It must be recognized that the notion of a *probability of a sentence* is an entirely useless one, under any interpretation of this term."

Noam Chomsky, 1969





Was ist statistische maschinelle Übersetzung?

▶ Gegeben:

- parallele Sätze von vorübersetztem Trainingsmaterial
- Beispiel: politische Reden im Europäischen Parlament,
 Bücher, die bereits in verschiedene Sprachen übersetzt wurden (z.B. die Bibel, Handbücher, Patente),

- - -

- ► Typische Größen (in Anzahl der parallelen Sätze):
 - ▶ klein: 40 K (IWSLT)



► Ziel:

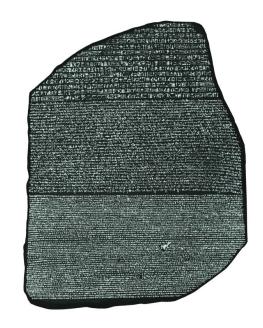
▶ Die beste (d.h.: wahrscheinlichste) Übersetzung eines unbekannten Satzes zu finden



Statistischer Ansatz

- ► Idee: die Übersetzung als Entschlüsselungsproblem zu sehen
- lacksquare Für einen gegebenen Quellsatz $f_1^J=f_1\dots f_j\dots f_J$
 - imes berechne $Pr(e_1^I|f_1^J)$ für alle möglichen Zielsätze $e_1^I=e_1\dots e_i\dots e_I$
 - ightharpoonup wähle den Satz $\hat{e}_1^{\hat{I}}$, der die Wahrscheinlichkeit maximiert

$$egin{aligned} \hat{e}_1^{\hat{I}} &= rg\max_{e_1^{I}} \left\{ p(e_1^{I}|f_1^{J})
ight\} \ &= rg\max_{e_1^{I}} \left\{ p(e_1^{I}) \cdot p(f_1^{J}|e_1^{I})
ight\} \end{aligned}$$





Charakteristika der maschinellen Übersetzung

- ► Training in zwei Phasen: Wort-Alignment und Phrasenextraktion
- **▶** Verschiedene Modelle:
 - ▶ Übersetzungsmodell
 - Sprachmodell
 - > Umordnungsmodell
 - ▶ Einige weitere zusätzliche Modelle...

Alle werden innerhalb eines statistischen Frameworks kombiniert

- ► Im Prinzip sollten alle Sätze der Zielsprache als mögliche Übersetzungen in Betracht gezogen werden
 - ▶ NP-vollständiges Problem
 - ▶ Effiziente Suchalgorithmen erforderlich
 - ⊳ Einschränkungen des Suchraums





Bestandteile der maschinellen Übersetzung

- ► Fünf integrale Bestandteile:
 - \triangleright Training: Wort-Alignment für $f_j \# e_i$ Wortpaare

 - ▶ Log-lineare Modellkombination: Kombination von Einzelmodellen, z.B. Übersetzungsmodell, Sprachmodell, ...
 - Suche: Finden des wahrscheinlichsten, "plausibelsten" Zielsatzes
 - Optimierung: automatische Bewertung der Ausgabe und Optimierung
- ► Ähnlichkeiten zur Spracherkennung, aber
 - ▶ Nicht-monotone Suche
 - ▶ Ausgabe schwerer zu interpretieren (z.B. bei Synonymen, anderer Grammatikstruktur, ...)

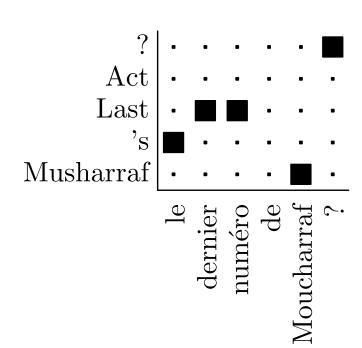




Training: Wort-Alignment

- ► Eingeführt von IBM 1989–1993
 - **⊳** Sequenz von IBM-1, ..., IBM-5 Modellen
 - Berechnung durch den EM-Algorithmus (iteratives Verfahren)
- Erweiterungen der RWTH

 - Effiziente Implementierung
 - ▶ Open Source Toolkit: GIZA++





Alignment-Modelle

$$egin{aligned} Pr(f_1^J|e_1^I) &= p(J|I) \cdot Pr(f_1^J|J,e_1^I) \ &= p(J|I) \cdot \sum_{a_1^J} Pr(f_1^J,a_1^J|J,e_1^I) \ &= p(J|I) \cdot \sum_{a_1^J} \prod_{j=1}^J p(a_j|j,J,I) \cdot p(f_j|e_{a_j}) \end{aligned}$$



Annahmen der Zero-Order Alignment-Modelle

$$Pr(f_1^J|e_1^I) = p(J|I) \cdot \sum_{a_1^J} \prod_{j=1}^J p(a_j|j,J,I) \cdot p(f_j|e_{a_j})$$

► Längenmodell: Abhängigkeit der Länge J nur vom Zielsatz e_1^I :

$$Pr(J|e_1^I) = p(J|I)$$

► Alignment-Modell: Abhängigkeit nur an der absoluten Position j (und den Längen J und I):

$$Pr(a_{j}|a_{1}^{j-1},J,e_{1}^{I})=p(a_{j}|j,J,I)$$

Lexikon Wahrscheinlichkeit: Abhängigkeit nur von e_i in der Position $i=a_j$:

$$Pr(f_j|f_1^{j-1},a_1^J,J,e_1^I) = p(f_j|e_{a_j})$$



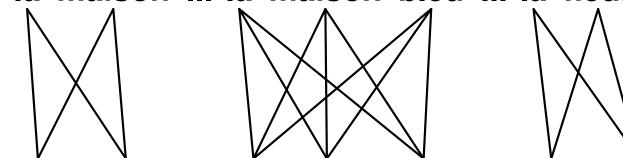


... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...



... la maison ... la maison bleu ... la fleur ...



... the house ... the blue house ... the flower ...



... la maison ... la maison bleu ... la fleur ...

the house ... the blue house ... the flower ...



... la maison ... la maison bleu ... la fleur ...

the house ... the blue house ... the flower ...



... la maison ... la maison bleu ... la fleur ...

the house ... the blue house ... the flower ...



Übung 2

- ► Ziel: Konstruktion eines (einfachen) Übersetzers für Französisch-Englisch
 - ▶ wortweise Übersetzung
 - > monoton von links nach rechts, keine Umordnungen in der Wortabfolge
 - \triangleright Zielsatz hat gleiche Länge wie Quellsatz, d.h. I=J.
 - ▶ kein Sprachmodell
- ► Eingabe: unbekannte Sätze in Französisch
- ► Ausgabe: Übersetzung in Englisch





这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	
the	7 people	including	by some		and	the russian	the	the astronauts		104
it	7 people inc	THE PROPERTY OF THE PROPERTY O	by france		and the	the russian	()	international astronautical	of rapporteur .	100
this	7 out	including the	from	the french	and the	russian	the fift	h		
these	7 among	including from		the french a	and	of the russian	of	space	members	· (4)
that	7 persons	including from	the	of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france ar	nd	russian		astronauts		. the
	7 numbers i	nclude	from france		and russ	ian	of astr	onauts who	. 2	
- 1	7 population	ns include	those from fran	ice and russian		Lack Court Rivers	astronauts.		dess	
10.	7 deportees	included	come from	france and russia		ssia	in	astronautical	personnel	
- 8	7 philtrum	including thos	e from	france an	d	russia	a space	e	member	
		including repr	esentatives from	france and	the	russia	()	astronaut	3	
		include	came from	france an	d russia		by cost	monauts		
		include represe	entatives from	french and russia		0.00	cosmonauts			
		include	came from fran	ce	and russ	ia 's	cosmonauts.			
		includes	coming from	french and	100000000000000000000000000000000000000	russia 's	(7)	cosmonaut	02	1
			36	french and	russian	Land Control of the C	's	astronavigation	member .	1
				french	and ru	ssia	astro	nauts	A second second	ł
- 3		Ç			and russ	ia 's			special rapporteur	
					, and	russia			rapporteur	t
					, and rus	sia		7	rapporteur.	1
		Î			, and rus	sia				
		i i			or	russia 's				1



这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	
the	7 people	including	by some		and	the russian	the	the astronauts		•
it	7 people inc	ruded	by france		and the	the russian		international astronautical	of rapporteur .	
thic	7 cat	including the	from	the french	and the	russian	the fift	h	540	
these	7 among	including from		the french a	and	of the russian	of	space	members	1.6
unat	7 persons	including from	the	of france	and to	russian	of the	aerospace	members .	1
	7 include		from the	of france ar	nd	russian	70	astronauts		. the
	7 numbers i	ndude	from france		and russi	an	of astr	stronauts who		. 27
	7 population	ns include	those from fran	ce	and russian			astronauts.		- Inches
15.	7 deportees	included	come from	france	and ru	ssia	in	astronautical	personnel	i
	7 philtrum	in luding thos	e from	france an	d	russia	a space	9	member	1
		including repr	esentatives from	france and	the	russia	Č.	astronaut	3	Ï
		include	came from	france an	d russia		by cost	monauts		1
		merade repres	ntatives from	french and rus		russia		cosmonauts		
		include	came from fran	ce	and russi	ia 's		cosmonauts.		00
		includes	coming from	french and		russia 's	cosmonaut		52	1
				french and	russian	***************************************	's	astronavigation	member .	
				french	and ru	ssia	astro	nauts		2
- 6		Ų.			and russi	ia 's			special rapporteur	
		Ť			, and	russia			rapporteur	1
					, and rus	sia		î	rapporteur.	W
ĺ		Ĵ			, and rus	sia		Ĺ		
					or	russia 's				



这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	
the	7 people	including	by some		and	the russian	the	the astronauts		
it	/ people inc	ruded	by france		and the	the russian		international astronautical	of rapporteur .	000
thic	7 cat	including the	from	the french	and the	russian	the fift	h	S21	
these	7 among	including from	Č.	the french a	and	of the russian	of	space	members	
tnat	7 persons	including from	the	of france	and to	russian	of the	201050200	mombore	*
	7 include		from the	of france ar	nd		7)	astronauts		. the
	7 numbers i	ndude	f om france		and russi	an	of astr	onauts who		. 27
	7 population	ns include	hose from fran	ce	and russian			astronauts .		
	7 deportees	included	come from	france	and rus	ssia	in	astronautical	personnel	ŝ
	7 philtrum	in luding thos		france an	d	russia	a space	e	member	1
		including repr	esentatives from	france and	the	russia	(I)	astronaut	51	1
		include	came from	france an	d russia	100	by cost	monauts		1
		merade repres	ntatives from	french and russia		ssia	cosmonauts			1
		include	came from fran	ce	and russi	a 's		cosmonauts.		077
		includes	coming from	french and		russia 's	Ű/	cosmonaut	50.	
				french and	russian	***************************************	's	astronavigation	member .	
1				french	and ru	ssia	astro	nauts		1
					and russi	a 's			special rapporteur	
					, and	russia			rapporteur	Ť
					, and rus	sia			rapporteur.	1//.
					, and rus	sia	0			1
					or	russia 's				

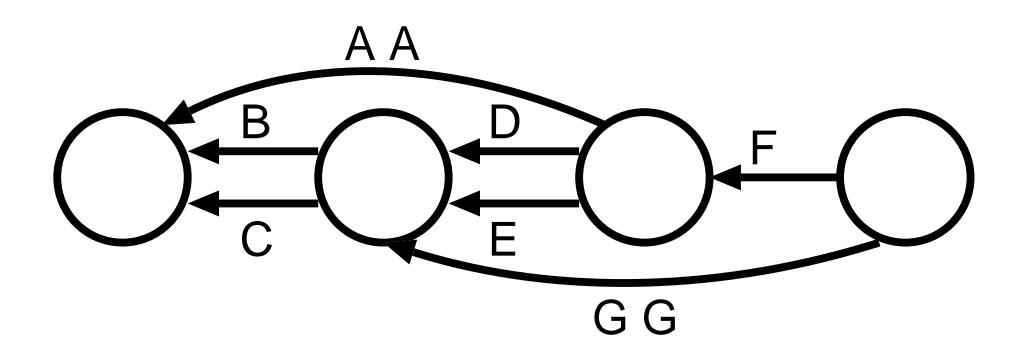


这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	
the	7 people	including	by some		and	the russian	the	the astronauts		
it	/ people inc	1172/05/05/05/05/05/05/05/05/05/05/05/05/05/	by france	_	and the	the russian	e entre e e e	international astronautical	of rapporteur .	100
thic	7 rat	including the		the french	and he	"Serff State Supposition of a	the fift		2	
these	7 among	including from		the french		of the russian	of	space	members	18
tnat	7 persons	including from		of france	and to	russian	of the	aemenana	members	
	7 include	37.	from the	of france ar	nd	The second		astronauts		. the
	7 numbers is	n lude	f om france		and russ	ian	of astr	onauts who		
	7 population	is include	hose from fran	co	and russ	ian	LICENSES STATES	astronauts.		
	7 deportees	included	come from	france	and ru	ssia	in	astronautical	personne)	;
	7 philtrum	in luding those	e from	папсе ап	d	russia	a space	9	member	
		including repr	esentatives from	france and	the	russia	()	act conditi	3	
		include	came from	f ance an	d russia		by cost	nonauts		
		merade represe	ntatives from	rrench and russia		Cocces orients				
		include	came from fran	ce	and russ	ia 's	cosmonauts.			· ·
		includes	coming from	french and		russia 's	Úr.	cosmonaut	52.]
				rench and	russian		's	astronavigation	member .	
				french	and ru	ssia	astro	nauts		1
		ï			and russ	ia 's			special rapporteur	
					, and	russia			rapporteur	
					, and rus	ssia			rapporteur.	W
					, and rus	ssia	0			
					or	russia 's				



Einzelwortbasierte monotone Übersetzung

- ► Einfach: Berechnung des Satzes mit den minimalen Kosten
- Problematisch: Berechnung der nächstbesten Alternative für beliebig lange Übergänge

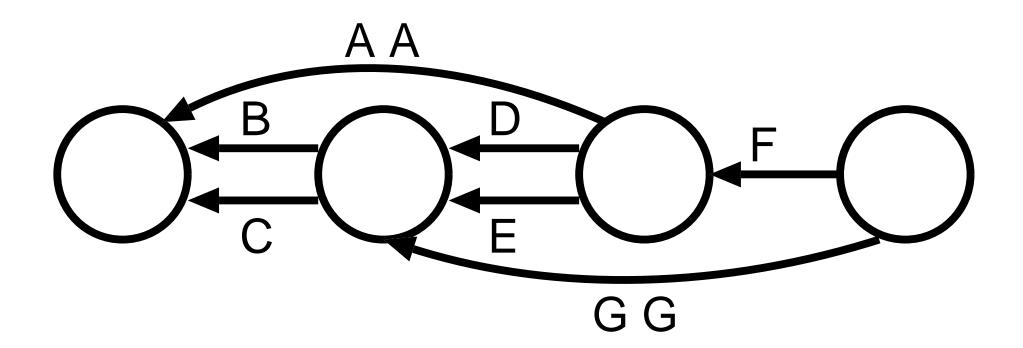




Einzelwortbasierte monotone Übersetzung

► Einfach: (langweilig)

► Problematisch: Eure Aufgabe (mit dem A*-Algorithmus)



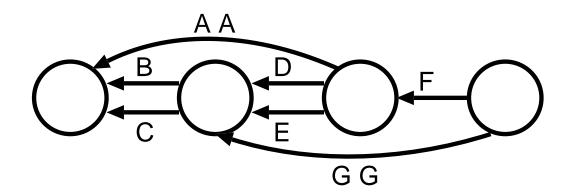


A*-Suche

- ► single-best Berechnung einfach: jeweils besten Pfad abspeichern
- ▶ n-best Berechnung durch den A*-Suchalgorithmus
 - ▶ informierter Suchalgorithmus
 - untersucht Knoten zuerst, die am vielversprechendsten sind
 - \triangleright benötigt *optimistische* Schätzungsfunktion f(x)
- lacksquare in unserer Übersetzung: f(x)=g(x)+h(x), mit
 - hd g(x) sind die Übersetzungskosten
 - $\triangleright h(x)$ sind die besten Pfade zum Knoten

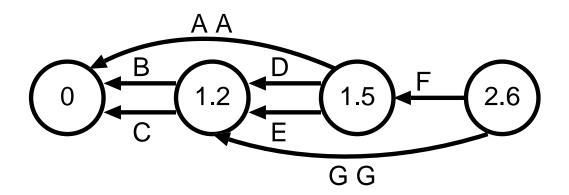


1.5 # X Y # A A
1.4 # X # B
1.2 # X # C
Tabelle: 1.2 # Y # D
1.3 # Y # E
1.5 # Y Z # G G
1.1 # Z # F





1.5 # X Y # A A
1.4 # X # B
1.2 # X # C
Tabelle: 1.2 # Y # D
1.3 # Y # E
1.5 # Y Z # G G
1.1 # Z # F

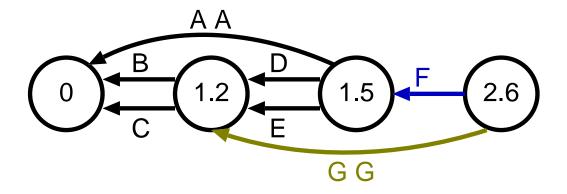




1.5 # X Y # A A
1.4 # X # B
1.2 # X # C
Tabelle: 1.2 # Y # D
1.3 # Y # E
1.5 # Y Z # G G
1.1 # Z # F

hyp	f(x)	g(x)	h(x)
F	2.6	1.1	1.5
G G	2.7	1.5	1.2

⇒ expandiere F



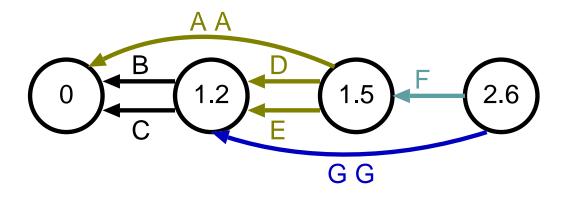


1.5	#	XY	#	AA
1.4	#	X	#	В
1.2	#	X	#	C
1.2	#	Y	#	D
1.3	#	Y	#	Ε
1.5	#	ΥZ	#	G G
1.1	#	Z	#	F

Tabelle:

hyp	f(x)	g(x)	h(x)
AAF	2.6	2.6	0
GG	2.7	1.5	1.2
DF	3.5	2.3	1.2
EF	3.6	2.4	1.2

- \Rightarrow gib A A F aus
- \Rightarrow expandiere G G



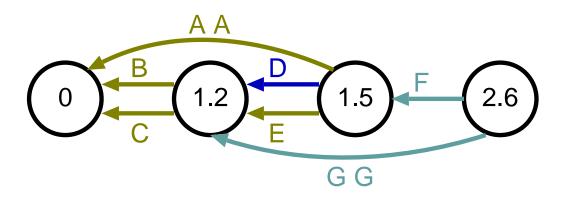


1.5	Ħ	XY	#	AA
1.4	#	X	#	В
1.2	#	X	#	C
1.2	#	Y	#	D
1.5	#	ΥZ	#	G G
1.1	#	Z	#	F
	1.4 1.2 1.2 1.3 1.5	1.4 # 1.2 # 1.2 # 1.3 # 1.5 #	1.4 # X 1.2 # X 1.2 # Y 1.3 # Y 1.5 # Y Z	1.5 # X Y # 1.4 # X # 1.2 # X # 1.2 # Y # 1.3 # Y # 1.5 # Y Z # 1.1 # Z #

VV

hyp	f(x)	g(x)	h(x)
CGG	2.7	2.7	0
BGG	2.9	2.9	0
DF	3.5	2.3	1.2
EF	3.6	2.4	1.2

- \Rightarrow gib C G G aus
- \Rightarrow gib B G G aus
- \Rightarrow expandiere D F





Fragen?

Ihr wisst wo unser Büro ist :-)

