

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)
- ▷ mittel: 3 M (Europarl)
- ▷ groß: 10 M (GALE)



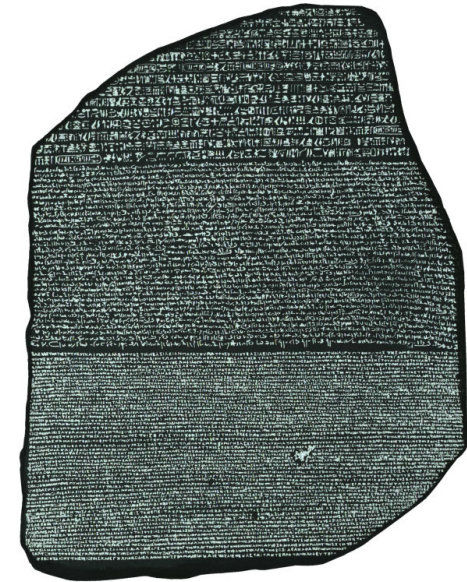
► Ziel:

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

Statistischer Ansatz

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

$$\begin{aligned}\hat{e}_1^I &= \arg \max_{e_1^I} \left\{ p(e_1^I | f_1^J) \right\} \\ &= \arg \max_{e_1^I} \left\{ p(e_1^I) \cdot p(f_1^J | e_1^I) \right\}\end{aligned}$$



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:

- ▷ **Training:** Wort-Alignment für $f_j \# e_i$ Wortpaare
- ▷ **Extraktion:** Extraktion von Fragmenten mit Wahrscheinlichkeiten aus einem bilingualen Trainings-Korpus
- ▷ **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**
 - ▷ **Zusätzliches Modell (HMM), statt IBM-2**
 - ▷ **Effiziente Implementierung**
 - ▷ **Open Source Toolkit: GIZA++**

?	■
Act
Last	.	■	■	.	.	.
's	■
Musharra	■	.
	le	dernier	numéro	de	Moucharra	?

Alignment-Modelle

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

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 $i = a_j$:

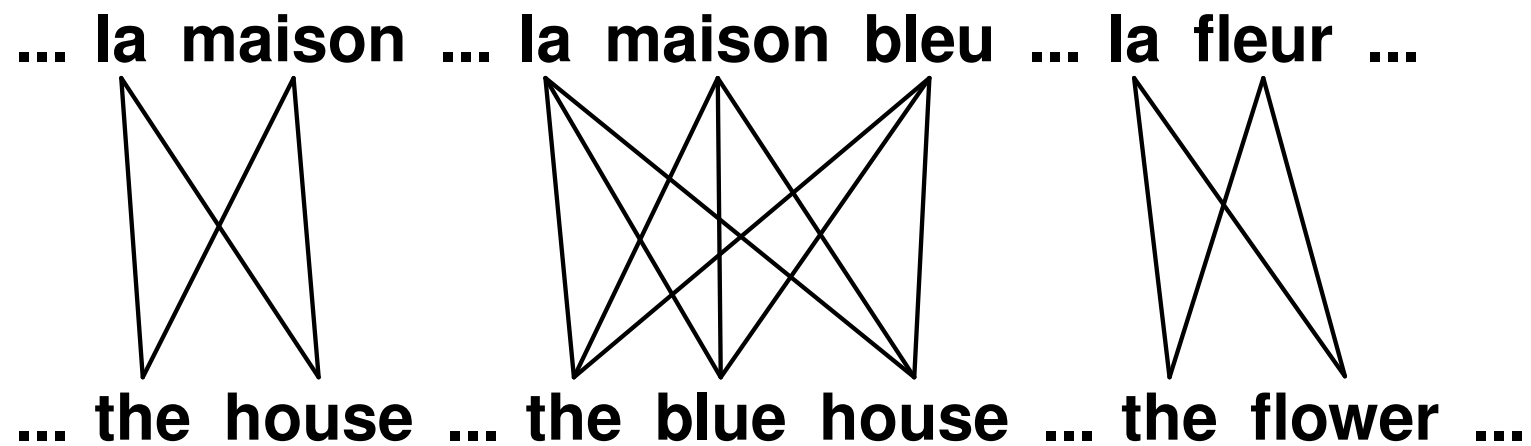
$$Pr(f_j | f_1^{j-1}, a_1^J, J, e_1^I) = p(f_j | e_{a_j})$$

Alignment-Training: EM Algorithmus

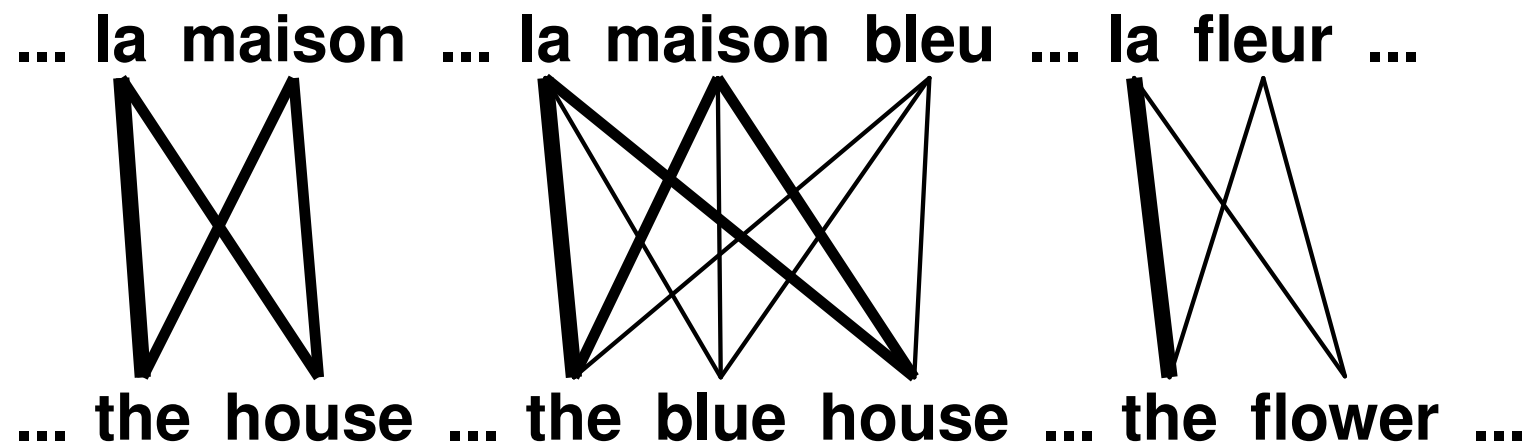
... la maison ... la maison bleu ... la fleur ...

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

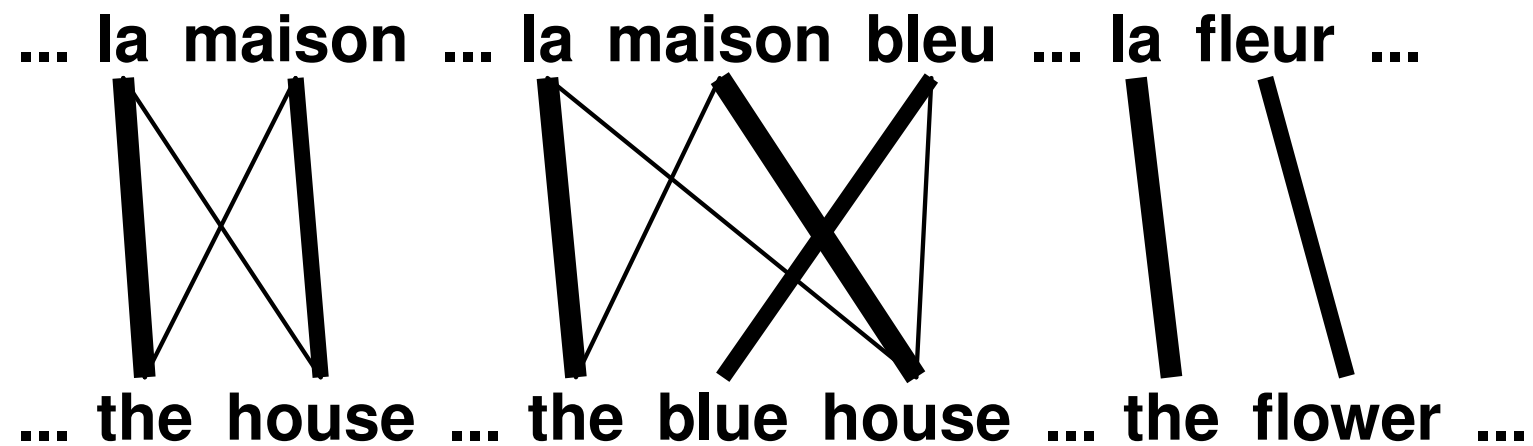
Alignment-Training: EM Algorithmus



Alignment-Training: EM Algorithmus

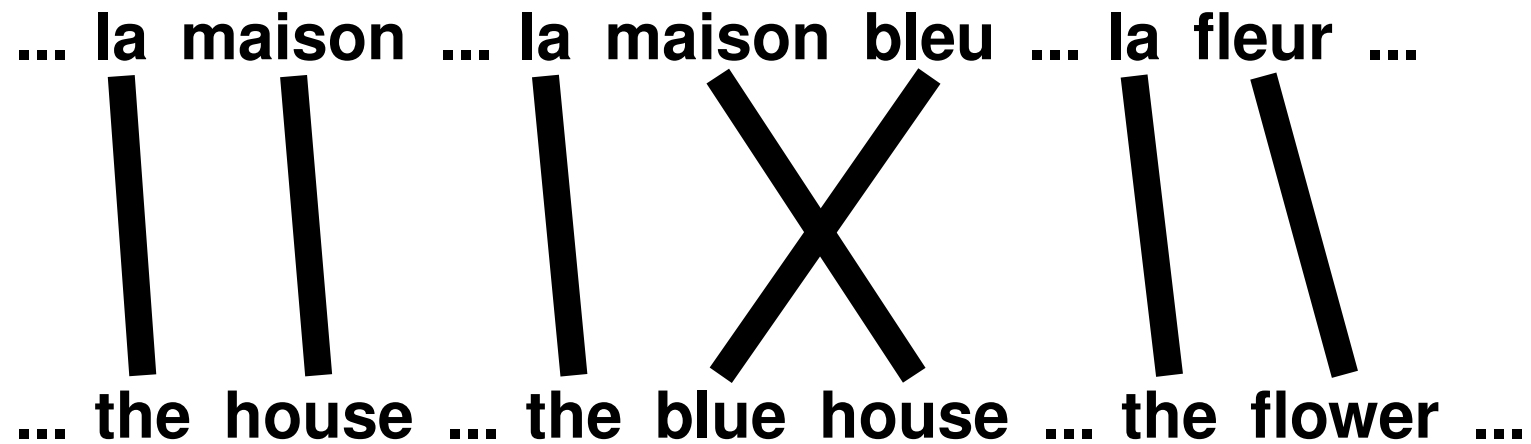


Alignment-Training: EM Algorithmus



Alignment-Training: EM Algorithmus

... 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
 - ▷ Zielsatz hat gleiche Länge wie Quellsatz, d.h. $I = J$.
 - ▷ kein Sprachmodell
- ▶ **Eingabe: unbekannte Sätze in Französisch**
- ▶ **Ausgabe: Übersetzung in Englisch**

Phrasenbasierte Übersetzung mit Umordnungen (Beispiel von Kevin Knight)

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 included		by france	and the	the russian		international astronautical	of rapporteur .
this	7 out	including the	from	the french	and the russian	the fifth		.
these	7 among	including from		the french and	of the russian	of	space	members .
that	7 persons	including from the		of france	and to	russian	of the	aerospace
	7 include		from the	of france and	russian		astronauts	. the
	7 numbers include		from france		and russian		of astronauts who	. "
	7 populations include		those from france		and russian		astronauts .	
	7 deportees included		come from	france	and russia	in	astronautical	personnel ;
	7 philtrum	including those from		france and	russia	a space		member
		including representatives from		france and the	russia		astronaut	
		include	came from	france and russia		by cosmonauts		
		include representatives from		french	and russia		cosmonauts	
		include	came from france		and russia 's		cosmonauts .	
		includes	coming from	french and	russia 's		cosmonaut	
				french and russian		's	astronavigation	member .
				french	and russia		astronauts	
					and russia 's			special rapporteur
					, and	russia		rapporteur
					, and russia			rapporteur .
					, and russia			
					or	russia 's		

Phrasenbasierte Übersetzung mit Umordnungen (Beispiel von Kevin Knight)

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 included	by france		and the	the russian		international astronautical	of rapporteur .
this	7 out	including the	from	the french	and the russian	the fifth		.
these	7 among	including from		the french and	of the russian	of	space	members .
that	7 persons	including from the		of france	and to	russian	of the	aerospace
	7 include	from the		of france and	russian		astronauts	. the
	7 numbers include	from france		and russian		of astronauts who		."
	7 populations include	those from france		and russian		astronauts .		
	7 deportees included	come from	france	and russia		in	astronautical	personnel ;
	7 philtrum	including those from	france and	russia		a space		member
		including representatives from	france and the	russia		astronaut		
		include	came from	france and russia		by cosmonauts		
		include representatives from	french	and russia		cosmonauts		
		include	came from france	and russia 's		cosmonauts .		
		includes	coming from	french and	russia 's	cosmonaut		
				french and russian	's	astronavigation	member .	
				french	and russia	astronauts		
				and russia 's			special rapporteur	
				, and	russia		rapporteur	
				, and russia			rapporteur .	
				, and russia				
				or	russia 's			

Phrasenbasierte Übersetzung mit Umordnungen (Beispiel von Kevin Knight)

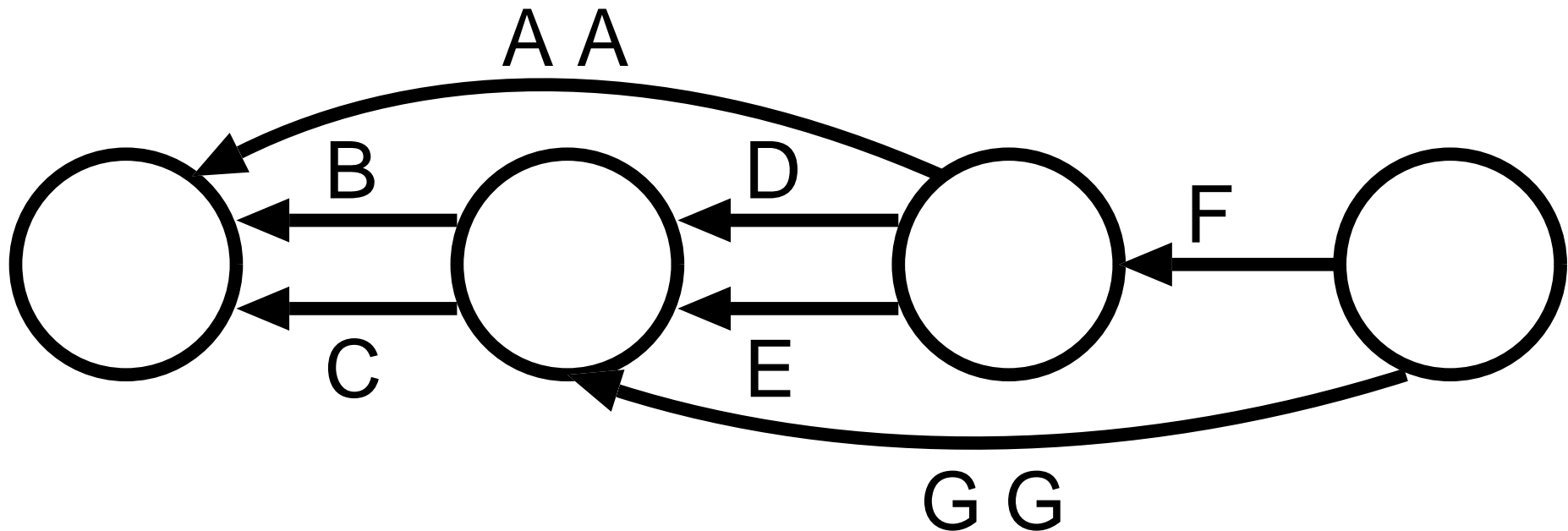
这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		and	the russian	the	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur	.
this	7 out	including the	from	the french	and the russian		the fifth			.
these	7 among	including from		the french and	of the russian	of	space	members		.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members	.
	7 include	from the		of france and	russian			astronauts	the	.
	7 numbers include	from france		and russian		of astronauts who				.
	7 populations include	those from france		and russian		astronauts				.
	7 deportees included	come from	france	and russia		in	astronautical	personnel		;
	7 philtrum	including those from	france and	russia		a space		member		.
		including representatives from	france and the	russia		astronaut				.
		include	came from	france and russia		by cosmonauts				.
		include representatives from	french	and russia		cosmonauts				.
		include	came from france	and russia 's		cosmonauts				.
		includes	coming from	french and	russia 's	cosmonaut				.
				french and russian	's	astronaut				.
				french	and russia	astronauts				.
				and russia 's				special rapporteur		.
				, and	russia			rapporteur		.
				, and russia				rapporteur		.
				, and russia						.
				or	russia 's					.

Phrasenbasierte Übersetzung mit Umordnungen (Beispiel von Kevin Knight)

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 included		by france		and the	the russian		international astronautical	of rapporteur	.
this	7 out	including the	from	the french	and the	russian	the fifth			.
these	7 among	including from		the french	and	of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members	.
	7 include		from the	of france and				astronauts		.
	7 numbers include		from france		and russian		of astronauts who			.
	7 populations include		those from france		and russian		astronauts			.
	7 deportees included		come from	france	and	russia	in	astronautical	personal	;
	7 philtrum	including those from		france and		russia	a space		member	.
		including representatives from		france and the		russia	astronaut			.
		include	came from	france and russia			by cosmonauts			.
		include representatives from		french	and	russia	cosmonauts			.
		include	came from france		and	russia 's	cosmonauts			.
		includes	coming from	french and		russia 's	cosmonaut			.
				french and russian		's	astronaut			.
				french	and	russia	astronauts			.
					and	russia 's			special rapporteur	.
						russia			rapporteur	.
						and			rapporteur	.
						and				.
						or	russia 's			.

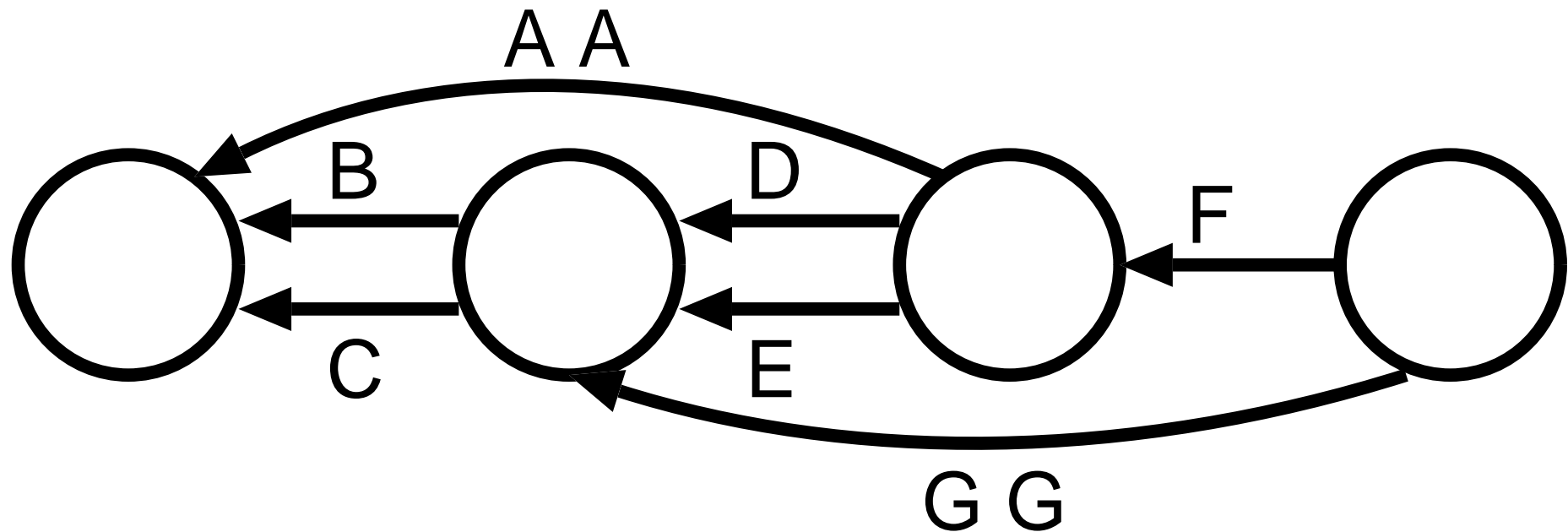
Einzelwortbasierte monotone Übersetzung

- ▶ **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)



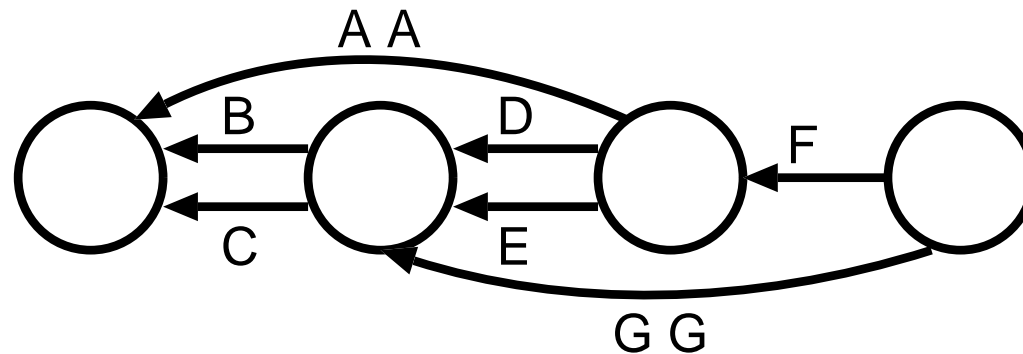
A*-Suche

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

N-Best Berechnung durch A*-Suche

Tabelle:

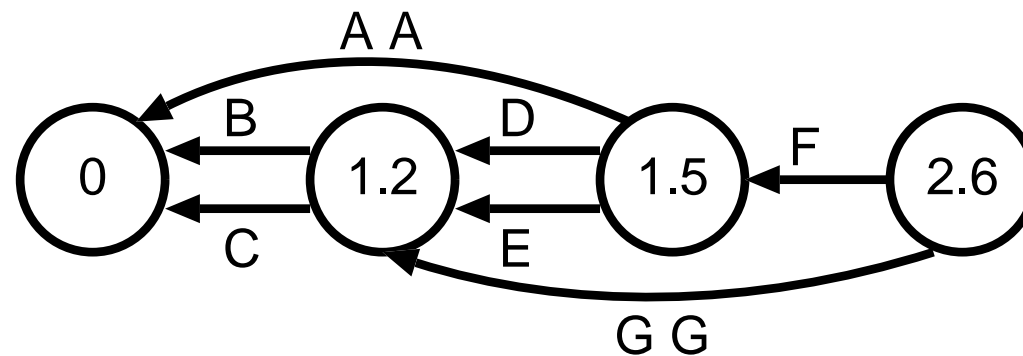
1.5	#	X	Y	#	A	A
1.4	#	X		#	B	
1.2	#	X		#	C	
1.2	#	Y		#	D	
1.3	#	Y		#	E	
1.5	#	Y	Z	#	G	G
1.1	#	Z		#	F	



N-Best Berechnung durch A*-Suche

Tabelle:

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



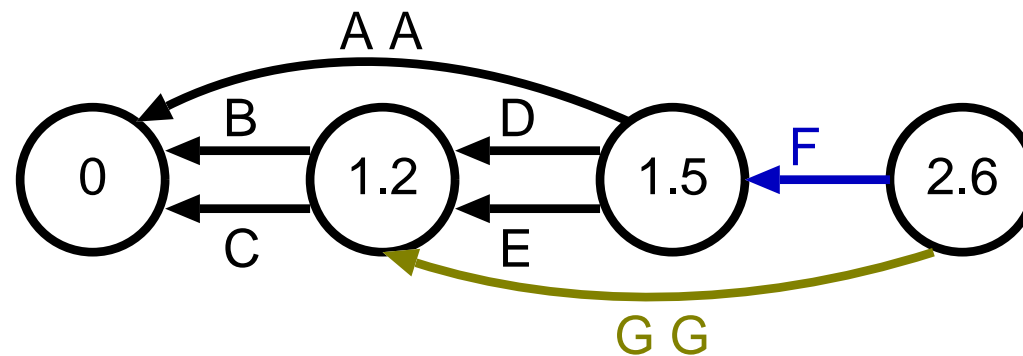
N-Best Berechnung durch A*-Suche

Tabelle:

1.5	#	X Y	#	A A
1.4	#	X	#	B
1.2	#	X	#	C
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



N-Best Berechnung durch A*-Suche

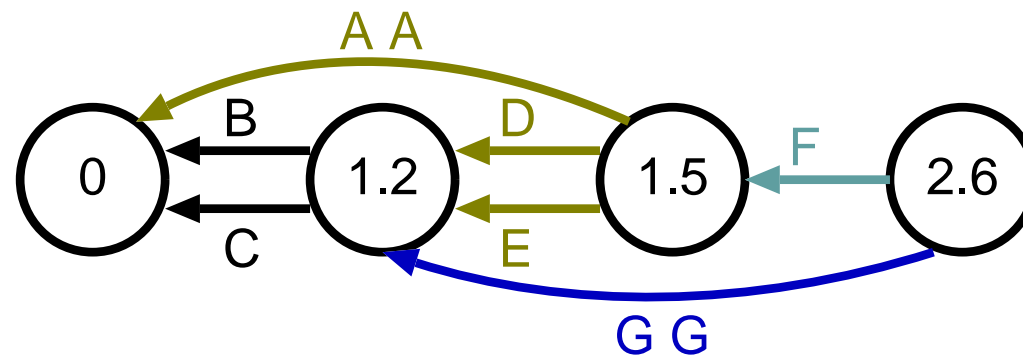
Tabelle:

1.5	#	X Y	#	A A
1.4	#	X	#	B
1.2	#	X	#	C
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)
A A F	2.6	2.6	0
G G	2.7	1.5	1.2
D F	3.5	2.3	1.2
E F	3.6	2.4	1.2

⇒ gib A A F aus

⇒ expandiere G G



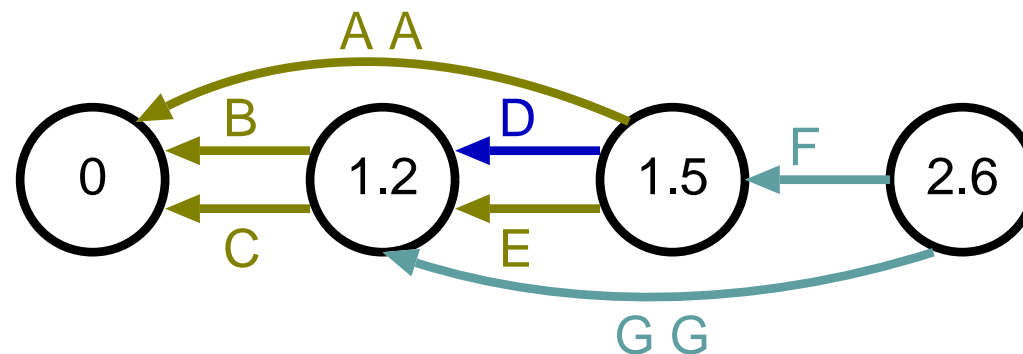
N-Best Berechnung durch A*-Suche

Tabelle:

1.5	#	X Y	#	A A
1.4	#	X	#	B
1.2	#	X	#	C
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)
C G G	2.7	2.7	0
B G G	2.9	2.9	0
D F	3.5	2.3	1.2
E F	3.6	2.4	1.2

- ⇒ gib C G G aus
- ⇒ gib B G G aus
- ⇒ expandiere D F



Fragen?

Ihr wisst wo unser Büro ist :-)

