



Diplomarbeit

Analyse und Synthese natürlichsprachiger Texte mit
rekurrenten neuronalen Netzen

- Vortrag -

Paul Werner

Hochschule für Technik und Wirtschaft Dresden



1. Zielstellung

- ▶ Analyse von natürlichsprachigen Sätzen und Ermittlung der Satzbedeutung unabhängig von der Formulierung.
- ▶ Algorithmische Erzeugung geeigneter Lern- und Validationsmengen (Beispielsätze) zum Anlernen und Testen des Netzes.
- ▶ Untersuchung zur Synthese (Erzeugung) von verschiedenartig formulierten aber inhaltlich synonymen Sätzen.



2. Beschreibung der Satzsemantik

- ▶ Zuordnung sogenannter semantischer Rollen für ausgewählte Satzglieder.
 - Max (Agens) liebt (Prädikat) Lisa (Patiens)
 - Lisa (Patiens) wird von Max (Agens) geliebt (Prädikat)
- ▶ Semantische Rollen sind keine grammatischen Satzglieder !
 - Subjekt, Prädikat, Objekt
- ▶ Rollenzuordnung bleibt gleich, trotz unterschiedlicher Formulierungen.
- ▶ Semantische Rollen können nach eigenem Ermessen aufgestellt werden.

Wort	sem. Rolle
Max	Agens
liebt	Prädikat
Lisa	Patiens



3. Mehrdeutigkeiten

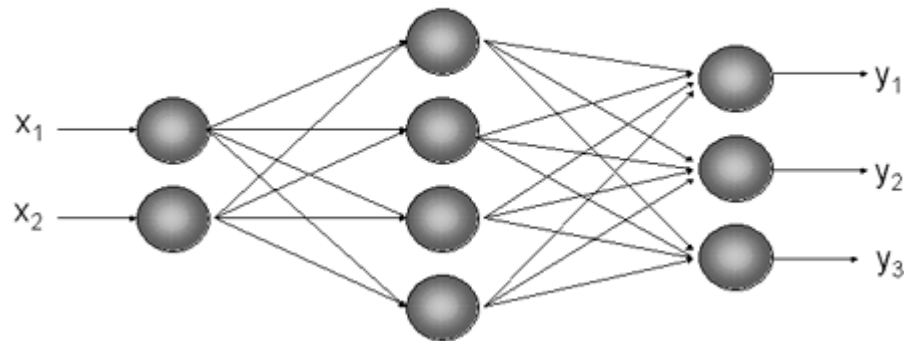
- ▶ Problem: Natürliche Sprache kann mehrdeutig formuliert werden.
(und das passiert häufig)
- ▶ Beispiel: “Time flies like an arrow”
 - Die Zeit vergeht so schnell, wie ein Pfeil vorüberfliegt.
 - Man soll Fliegen genauso timen, wie es ein Pfeil tut.
 - Man soll Fliegen genauso timen, die ähnlich wie ein Pfeil aussehen.
 - “Zeitfliegen” mögen einen Pfeil
- ▶ Syntaktisches Wissen alleine nicht ausreichend zur Ermittlung der Satzbedeutung.
- ▶ Kontextuelles Wissen ist notwendig !
- ▶ Komplettes “Weltwissen” kann in einem Parser nicht gespeichert und verarbeitet werden.



4. Neuronales Netz

- ▶ Idee: Motivation durch menschliche Sprachfähigkeit / Neurobiologie.
- ▶ Einen verhältnismäßig kleinen Anteil von natürlichsprachigen Satzbeispielen anlernen mit korrekt zugeordneten semantischen Rollen.
- ▶ Überwachtes Lernen, ähnlich wie der kindliche Spracherwerb.
- ▶ Abstraktion und Konzeptbildung (Generalisierung) zur korrekten Analyse nicht angelernter Sätze.
- ▶ Korrekte Zuordnung von semantischen Rollen der Satzglieder für nicht angelernte Sätze.

5. Feed-Forward-Netz

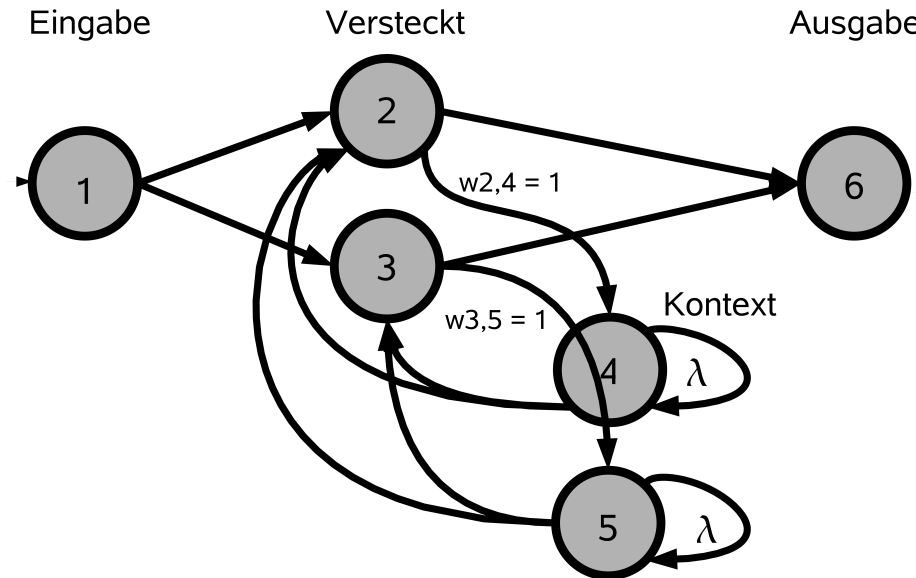


Einschränkungen:

- ▶ Eingabeschicht hat eine feste Größe (Anzahl Eingabeneuronen).
- ▶ Sätze unterschiedlicher Größe (Wortanzahl, Satzlänge) können nicht verarbeitet werden.
- ▶ Kein “internes Gedächtnis” => Dadurch können auch keine einzelnen Worte propagiert werden.
- ▶ Wenig Gemeinsamkeiten dem menschlichen neurobiologischen Sprachverstehen.

(Funktioniert so das Gehirn beim Hören und Lesen von Text?)

6. Elman-Netz



- ▶ “Internes Gedächtnis” durch Kontextneuronen (Nr. 4 und 5)
 - Speichern Aktivierung der versteckten Neuronen (Nr. 2 und 3) vom vorausgehenden Zeitschritt abhängig vom Gedächtnisfaktor $\lambda \in [0, 1]$.
- ▶ Ist in der Lage, zeitliche Abhängigkeiten von Eingaben zu verarbeiten.

Idee:

- ▶ Satz wird sequentiell (Wort für Wort) angelegt. (Auch beim Anlernen)
- ▶ Ausgabe der Worte mit ihren zugeordneten semantischen Rollen.



7. Lernmenge / Validationsmenge

- ▶ Definition einer generativen Grammatik zur Erzeugung der Lern- und Validationsmenge (Beispielsätze).
 - Nichtterminale werden ersetzt durch konkrete Wortbelegungen (s. Folie).
- ▶ Es entsteht ein großer Sprachumfang (Bsp: $12 \times 3 \times 3 \times 3 \times 3 \times 3 = 2916$ Sätze).
- ▶ Nur ein Bruchteil davon soll in die Lern- und Validationsmenge aufgenommen werden $\Rightarrow$ Lerndichte $\rho \in [0, 1]$.
 - Netz soll generalisieren. Eine verhältnismäßig kleine Lernmenge soll ausreichen.
- ▶ $\text{Validationsmenge} \subseteq \text{Gesamtmenge} \setminus \text{Lernmenge}$

8. Kodierung der Netzeingabe

► Beispiel: Aktivsatz “**Max (Agens) liebt (Prädikat) Lisa (Patiens)**”

■ Lernmuster

UND

■ Wörter- und Rollenbuch

	Zeitindex	t ₀	t ₁	t ₂
Eingabe:	Wort	Max	liebt	Lisa
Ausgabe:	Wort	Max	liebt	Lisa
(Soll-Werte)	sem. Rolle	Agens	Prädikat	Patiens

Wort	Code
Max	001
liebt	010
Lisa	100

Rolle	Code
Agens	001
Patiens	010
Prädikat	100

► Ein- und Ausgabe für das Netz

(Zeitindex)

Eingabe (Wort)	0 0 1
Ausgabe (Wort+Rolle)	0 0 1 0 0 1
Eingabe (Wort)	0 1 0
Ausgabe (Wort+Rolle)	0 1 0 1 0 0
Eingabe (Wort)	1 0 0
Ausgabe (Wort+Rolle)	1 0 0 0 1 0

t_0

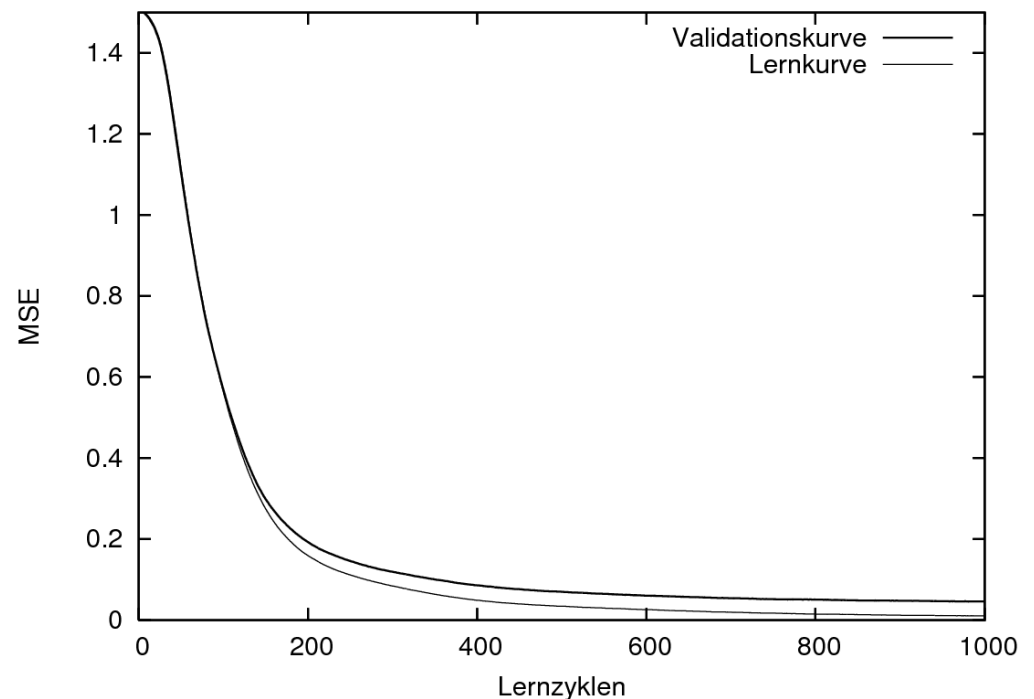
t_1

t_2

9. Lernprozess Aktiv- und Passivsätze

Beispiel Aktiv- und Passivsätze

- ▶ Gesamtmenge von 7488 Sätzen.
- ▶ Lerndichte $\rho = 0.01$ ergab 75 Sätze in der Lernmenge und 64 Sätze in der Validationsmenge (1% des gesamten Sprachumfanges).
- ▶ Gedächtnisfaktor $\lambda = 0.2$, Lernrate $\eta = 0.01$, Lernzyklen $n = 1000$.
- ▶ Elman-Netz
 - 2 versteckte Schichten
 - 50 Neuronen pro Schicht





10. Praxistest für Aktiv- und Passivsätze

- ▶ Praxistest mit allen 64 Sätzen (Aktiv- und Passivsätze gemischt) der Validationsmenge
 - Für alle Tests gilt: *Validationsmuster* $\neq$ *Lernmenge*.

semantische Rolle	Soll-Zuordnungen	Ist-Zuordnungen	%
Agens	64	58	90.6 %
Patiens	64	60	93.8 %
Prädikat	64	64	100.0 %
Ort	64	64	100.0 %
Zeit	64	64	100.0 %
Gesamt	320	310	96.9 %

- ▶ Variante Rollenbelegungen “sehr gut” gelernt und abstrahiert (> 90%).
- ▶ Invariante Rollenbelegungen “perfekt” gelernt, da bestimmte Worte immer mit einer bestimmten semantischen Rolle auftreten (gestern, heute, morgen).



12. Lernprozess mehrdeutige Sätze

- ▶ Motivation: Bestimmung der semantischen Rolle des letzten Satzteils
- ▶ Ausschließlich Sätze der folgenden Art anlernen:
 - AGENS fesselte PATIENS mit dem Seil (Instrument).
 - AGENS fesselte PATIENS mit dem Fernglas (Attribut).
 - AGENS beobachtet PATIENS mit dem Seil (Attribut).
 - AGENS beobachtet PATIENS mit dem Fernglas (Instrument).und
 - AGENS sieht PATIENS mit dem Fernglas (Instrument)
- ▶ Interessante Frage: Wie reagiert das Netz auf den nicht angelernten Satz
 - AGENS sieht PATIENS mit dem Seil (???)



12. Lernprozess mehrdeutige Sätze

▶ Parameter:

Gedächtnisfaktor $\lambda = 0.3$, Lernrate $\eta = 0.03$, Lernzyklen $n = 2000$.

▶ Elman-Netz

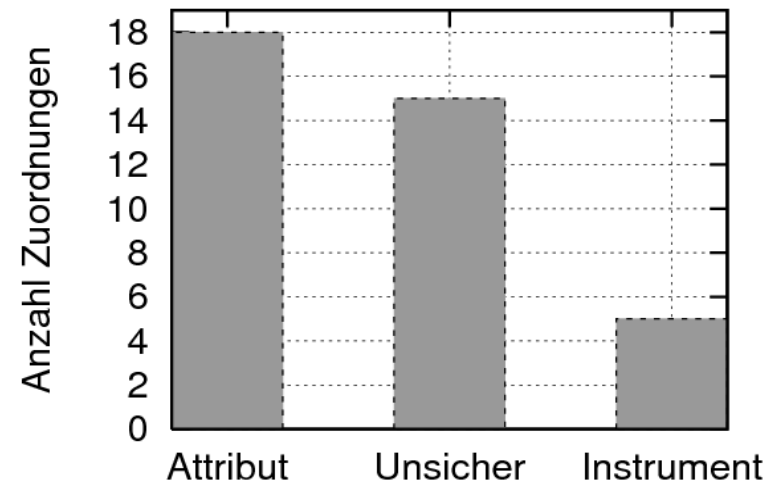
- 2 versteckte Schichten
- 50 Neuronen pro Schicht

13. Praxistest für mehrdeutige Sätze

- ▶ Praxistest ausschließlich mit Sätzen der Art (37 Sätze):
 - AGENS sieht PATIENS mit dem Seil. (nicht angelernt)
- ▶ Welche semantische Rolle wird dem Satzteil “mit dem Seil” zugeordnet ?
(Instrument oder Attribut ?)
- ▶ Ergebnis: Attribut: 18 Sätze, Unsicher: 14 Sätze, Instrument: 5 Sätze

Fazit:

- ▶ Das Netz hat den alternierenden Zusammenhang zwischen Instrument und Attribut gelernt (“Seil” und “Fernglas”).
- ▶ Das Wort “sieht” wurde nie im Zusammenhang mit Attribut angelernt.
- ▶ Bestimmte Worte verhalten sich nun wie Operatoren !





13. Praxistest für mehrdeutige Sätze

► Anschauliche Interpretation:

► **WENN**

- Max fesselte Lisa mit dem Seil => (Instrument).
- Max fesselte Lisa mit dem Fernglas => (Attribut).
- Max beobachtet Lisa mit dem Seil => (Attribut).
- Max beobachtet Lisa mit dem Fernglas => (Instrument).



13. Praxistest für mehrdeutige Sätze

► Anschauliche Interpretation:

► **WENN**

- Max fesselte Lisa mit dem Seil => (Instrument).
- Max fesselte Lisa mit dem Fernglas => (Attribut).
- Max beobachtet Lisa mit dem Seil => (Attribut).
- Max beobachtet Lisa mit dem Fernglas => (Instrument).

UND

- Max **X** Lisa mit dem Fernglas => (Instrument)



13. Praxistest für mehrdeutige Sätze

► Anschauliche Interpretation:

► **WENN**

- Max fesselte Lisa mit dem Seil ==> (Instrument).
- Max fesselte Lisa mit dem Fernglas ==> (Attribut).
- Max beobachtet Lisa mit dem Seil ==> (Attribut).
- Max beobachtet Lisa mit dem Fernglas ==> (Instrument).

UND

- Max **X** Lisa mit dem Fernglas ==> (Instrument)

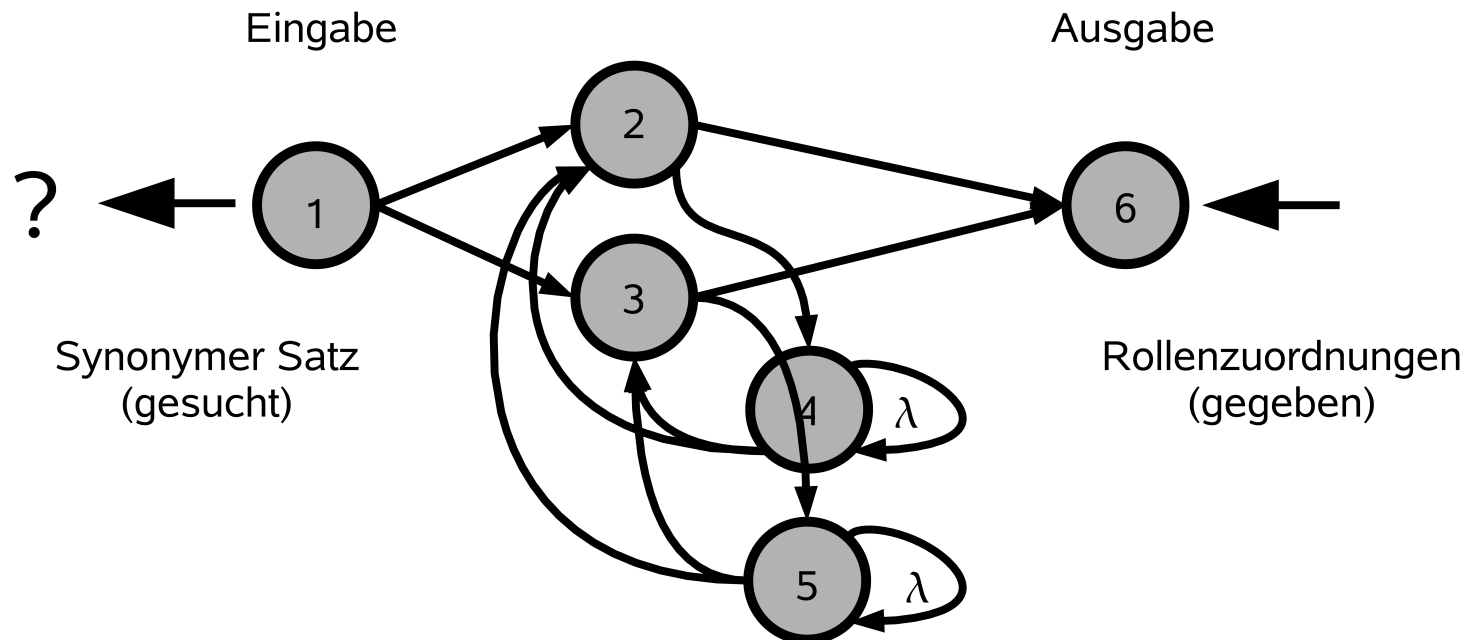
► **DANN**

- Max **X** Lisa mit dem Seil ==> (Attribut)

14. Synthese

Ziel:

- ▶ Erzeugung von synonymen Sätzen ausgehend von der Satzbedeutung.
 - Max sieht Lisa heute am Bahnhof.
 - Am Bahnhof wird Lisa heute von Max gesehen.
- ▶ Satzbedeutung in Form von Zuordnungen semantischer Rollen.





15. Kombinatorik

Idee:

- ▶ Alle möglichen Eingaben “probieren”.
- ▶ Beste Übereinstimmung der Ausgabe mit den geforderten Rollenzuordnungen finden $\Rightarrow$ bester synonymer Satz.

Problem:

- ▶ Berechnungskomplexität !

p^q Möglichkeiten mit

$p \dots$ Anzahl möglicher Worte

$q \dots$ maximale Satzlänge (Wortanzahl)



16. Genetischer Algorithmus

Bessere Idee: Genetischer Algorithmus !

- Ein möglicher Eingabesatz wird durch ein Individuum dargestellt.
- Ein Allel im Chromosomen entspricht einem Wort im Satz.
- ▶ Definition der Referenzausgabe (Soll-Zuordnung der semantischen Rollen)
- ▶ Bewertung und Vergleich der Ist-Ausgabe des Netzes mit der Soll-Ausgabe (Referenzausgabe) => Fitness-Funktion



17. Strukturen für Soll- und Ist-Ausgabe

► Referenzausgabe (Soll)

Wort	semantische Rolle		Referenzvektor $\vec{b} \in V$
0001 ("Lisa")	0100 ("Agens")	$\Rightarrow$	00010100
0100 ("liebt")	1000 ("Prädikat")		01001000
0010 ("heute")	0010 ("Zeit")		00100010
1000 ("Max")	0001 ("Patiens")		10000001

► Netz-Ausgabe (Ist)

Wort	semantische Rolle		Netzausgabevektor $\vec{a} \in U$
0.1 0.0 0.0 0.9	0.0 1.0 0.1 0.1	$\Rightarrow$	0.1 0.0 0.0 0.9 0.0 1.0 0.1 0.1
0.0 0.9 0.0 0.1	0.9 0.1 0.1 0.1		0.0 0.9 0.0 0.1 0.9 0.1 0.1 0.1
0.1 0.1 0.9 0.0	0.1 0.0 1.0 0.0		0.1 0.1 0.9 0.0 0.1 0.0 1.0 0.0
1.0 0.1 0.0 0.1	0.0 0.1 0.0 0.9		1.0 0.1 0.0 0.1 0.0 0.1 0.0 0.9



18. Fitness-Funktion

- ▶ Richtungskosinus $\cos(\vec{a}, \vec{b}) \in [0, 1]$ zwischen allen Ausgabevektoren und Referenvektoren berechnen.
 - Für Wort und Rolle: $\cos(\vec{a}, \vec{b}) \in [0, 2]$
- ▶ Summation aller Richtungskosinuse ergibt ersten Fitness-Wert f .

Negative Bewertungen (Minderung der Fitness):

- ▶ Bewertung falscher Netzausgaben.
 - (Erzeugt das Netz Ausgaben, welche in den Referenzwerten nicht vorkommen?)
 - Für jede falsche Netzausgabe, Fitness f mit dem Faktor $\alpha \in [0, 1]$ multiplizieren.

$$f = f \cdot \alpha^n \quad n \dots \text{Anzahl falscher Netzausgaben}$$



18. Fitness-Funktion

Negative Bewertungen (Minderung der Fitness):

► Bewertung mehrfacher Worte (Doubletten)

(“Lisa Lisa Lisa liebt Max”)

- $d \dots$ Anzahl mehrfacher Worte

- $f = f \cdot \beta^d \quad \beta \in [0, 1]$



19. Synthesevorgang

► Referenzwerte (Soll-Ausgaben)

Worte:	„Max“	„hört“	„Peter“	„heute“	„in der Stadt“
sem. Rolle:	Agens	Prädikat	Patiens	Zeit	Ort

► GA-Parameter: $n=30$, $c_{pop}=1000$, $\kappa=0.8$, $\mu=0.5$, $\alpha=0.9$, $\beta=0.5$
elite = 'survive', select = 'roulette', cross = 'uniform'

► Ergebnisse von 5 abgeschlossenen Evolutionen:

Nr.	Fitness	nach Zyklen	Propagierungen	erzeugter Satz
1)	8.9696	11	23.100	heute hört max der stadt peter
2)	8.9649	12	25.200	heute in stadt hört max peter
3)	8.9473	24	50.400	in max stadt peter hört heute
4)	8.9163	16	33.600	hört max peter heute der stadt
5)	8.0830	24	50.400	heute hört der stadt ruft max peter



19. Synthesevorgang

► Weitere Ergebnisse (verschiedene Referenzwerte):

Fitness	erzeugter Satz
8.9813	morgen ruft lisa moritz am bahnhof
9.9616	gestern fragt bahnhof peter lisa
8.9783	morgen der stadt hört max moritz

Problem:

- GA konvergiert stets gegen einen kurzen prägnanten Aktivsatz.
- Es werden keine Passivsätze gefunden.

Idee:

- Notwendige Worte (“wird”, “von”, “ge”) für Passivsätze in der richtigen Reihenfolge mit der Fitness-Funktion bewerten.
 - Worte und ihre Reihenfolge muss vorgegeben werden.



20. Synthesevorgang Passivsätze

- ▶ Vorgabe: (Max, wird, von, Lisa, ge)

$$\text{orderness} = \frac{c}{m}$$

m ... Gesamtanzahl der Worte vom Eingabesatz (Ist – Ausgabe)
 c ... Anzahl der vorhandenen und korrekt positionierten Worte
 $\text{orderness} \in [0, 1]$

- ▶ Anpassung Fitness-Wert

$$f = f \cdot \text{orderness}$$

- ▶ Ergebnisse

Nr.	Fitness	nach Zyklen	erzeugter Satz
1)	6.5419	13	heute wird lisa von max gehört der stadt
2)	6.4302	17	heute wird lisa gehört max der stadt
3)	6.4247	28	heute gehört wird der stadt lisa von max
4)	6.8134	27	wird heute lisa von max gehört
5)	5.2259	4	der stadt heute hört wird lisa von max
6)	5.1508	9	heute gehört wird lisa von max



Ende

Danke für Ihre Aufmerksamkeit!