Karlsruher Institut für Technologie

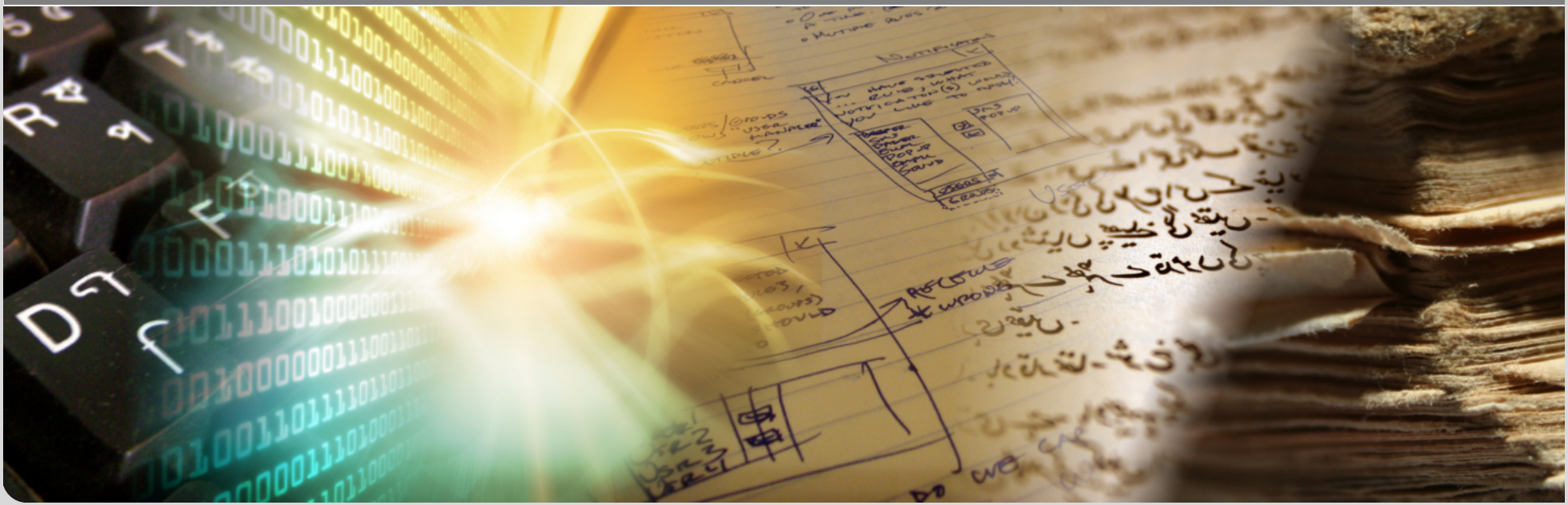
Fakultät für Informatik

Masterarbeit

Aufbereitung von Spracherkennerausgaben

Sven Scheu | 15. Juli 2016
Betreut von Sebastian Weigelt

IPD TICHY, FAKULTÄT FÜR INFORMATIK



Motivation

- Programmierung mit Sprache
- Bessere Spracherkennung
- $\rightarrow$ Einfachere und bessere Weiterverarbeitung

Spracherkenner

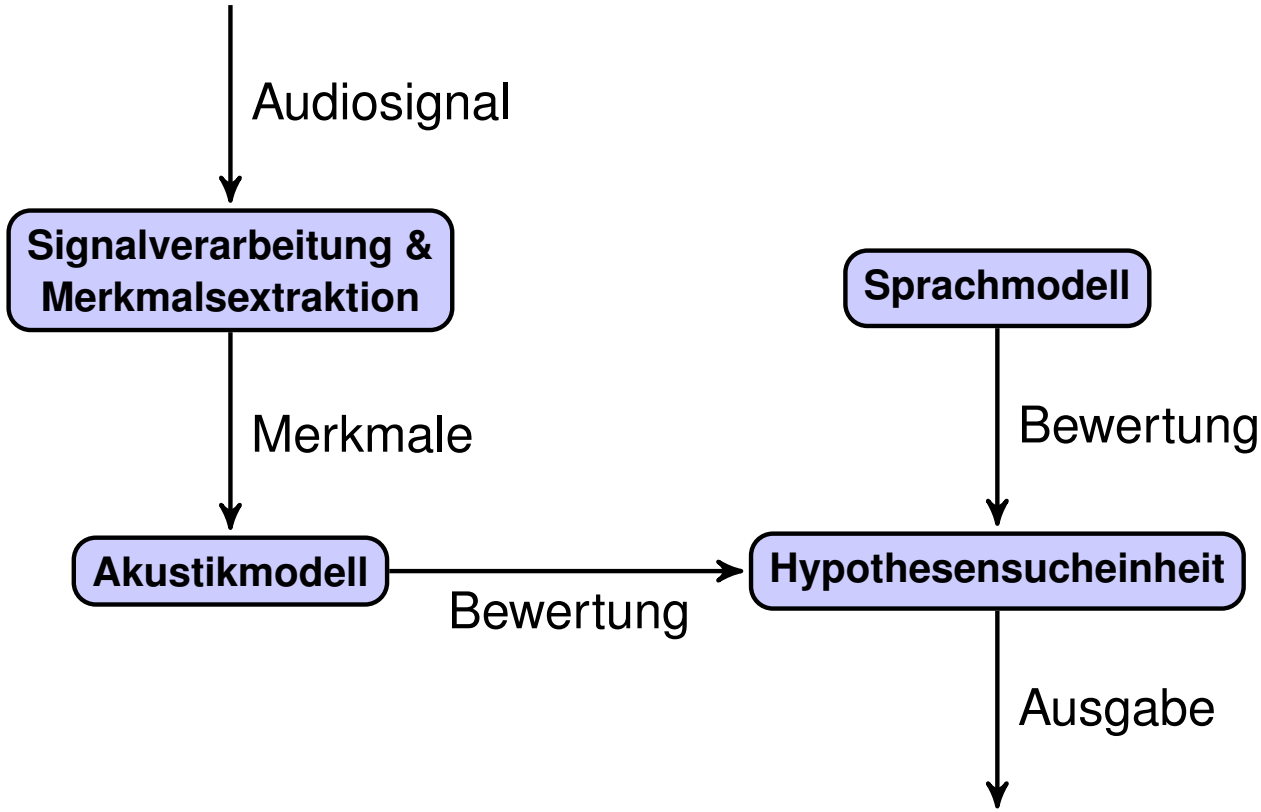


Abbildung: Spracherkenner Grundaufbau [YD14]

Wortfehlerrate

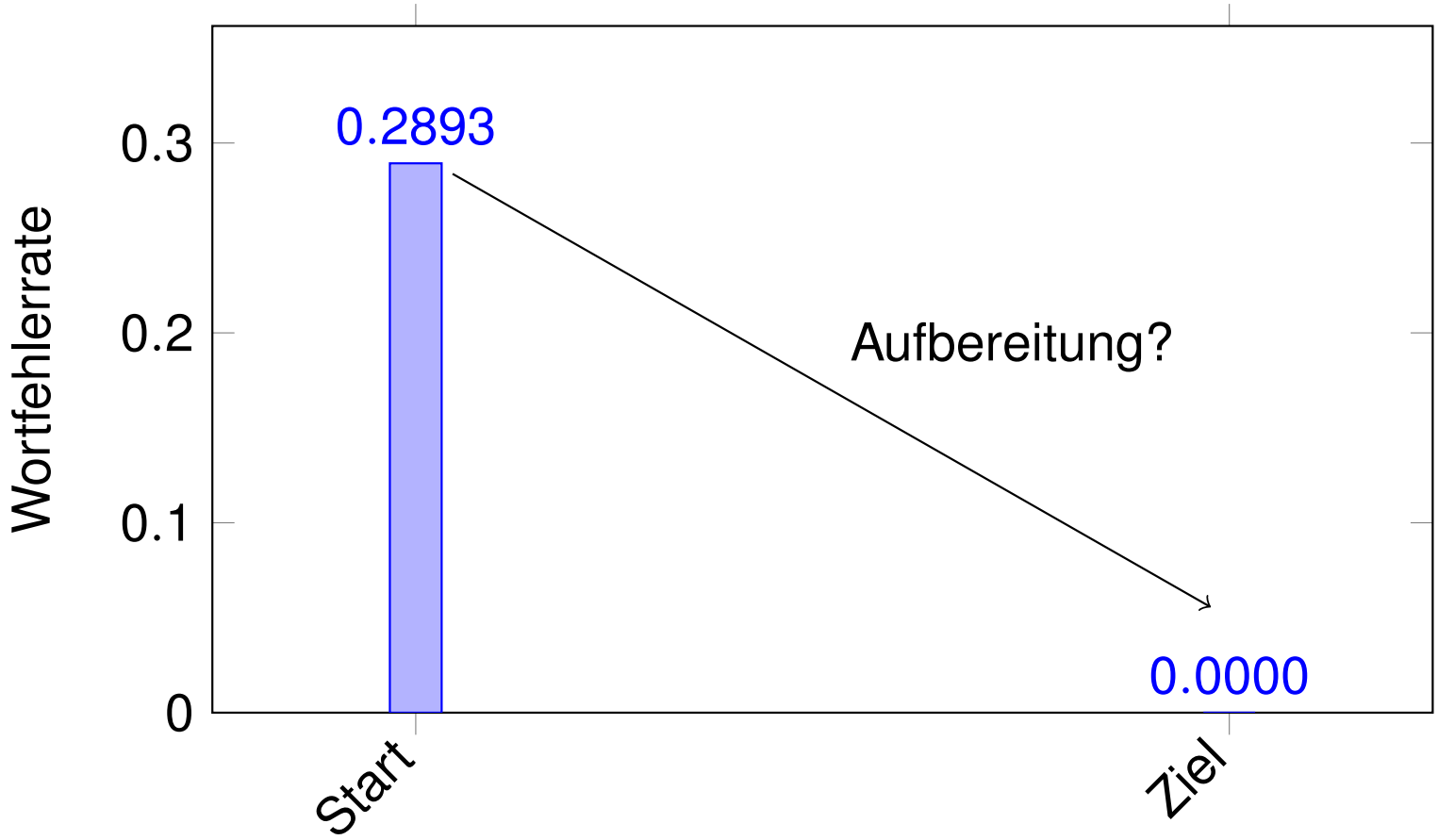
Wortfehlerrate

$$\frac{\textit{Einfügungen} + \textit{Substitutionen} + \textit{Löschungen}}{\textit{Wortzahl des Referenztexts}}$$

WordNet

- Datenbank der Princeton University [Pri10]
- Synsets
- Glosse

Ausgangspunkt und Ziel



Verwandte Arbeiten

Verwirrungsnetzwerke

- Kombination von Spracherkennern [Fis97]
- Kombination von Maschinenübersetzern [BBR01; MUN06]
- Reduktion der Wortfehlerrate statt der Satzfehlerrate [MBS00]

Websuche basierte *N*-Gramme

- Erzeugung neuer *N*-Gramme durch Websuche [NH05; KL03]

Verwandte Arbeiten

Verwirrungsnetzwerke

- Kombination von Spracherkennern [Fis97]
- Kombination von Maschinenübersetzern [BBR01; MUN06]
- Reduktion der Wortfehlerrate statt der Satzfehlerrate [MBS00]

Websuche basierte *N*-Gramme

- Erzeugung neuer *N*-Gramme durch Websuche [NH05; KL03]

Verwirrungsnetzwerke

- Kombination mehrerer Ausgaben auch Spracherkennerübergreifend
- Kompakte Darstellung der Spracherkennerausgaben

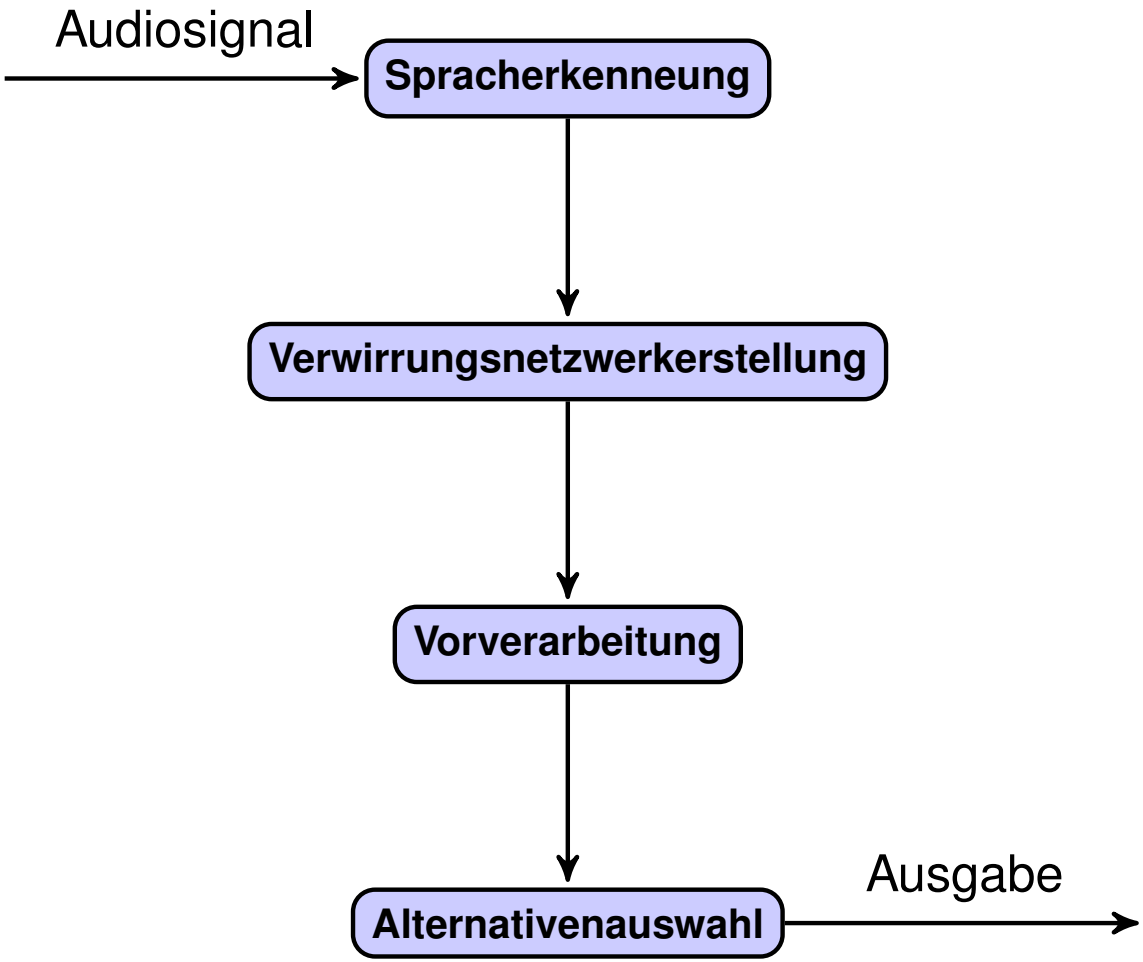
- Eigenschaften
  - Gerichtet
  - Beschriftet
  - Zusammenhängend
  - Kreisfrei
  - Alle Pfade starten an einem Quellknoten.
  - Alle Pfade enden in einem Senkenknoten.
  - Alle Pfade gehen durch alle Knoten.

Verwirrungsnetzwerke

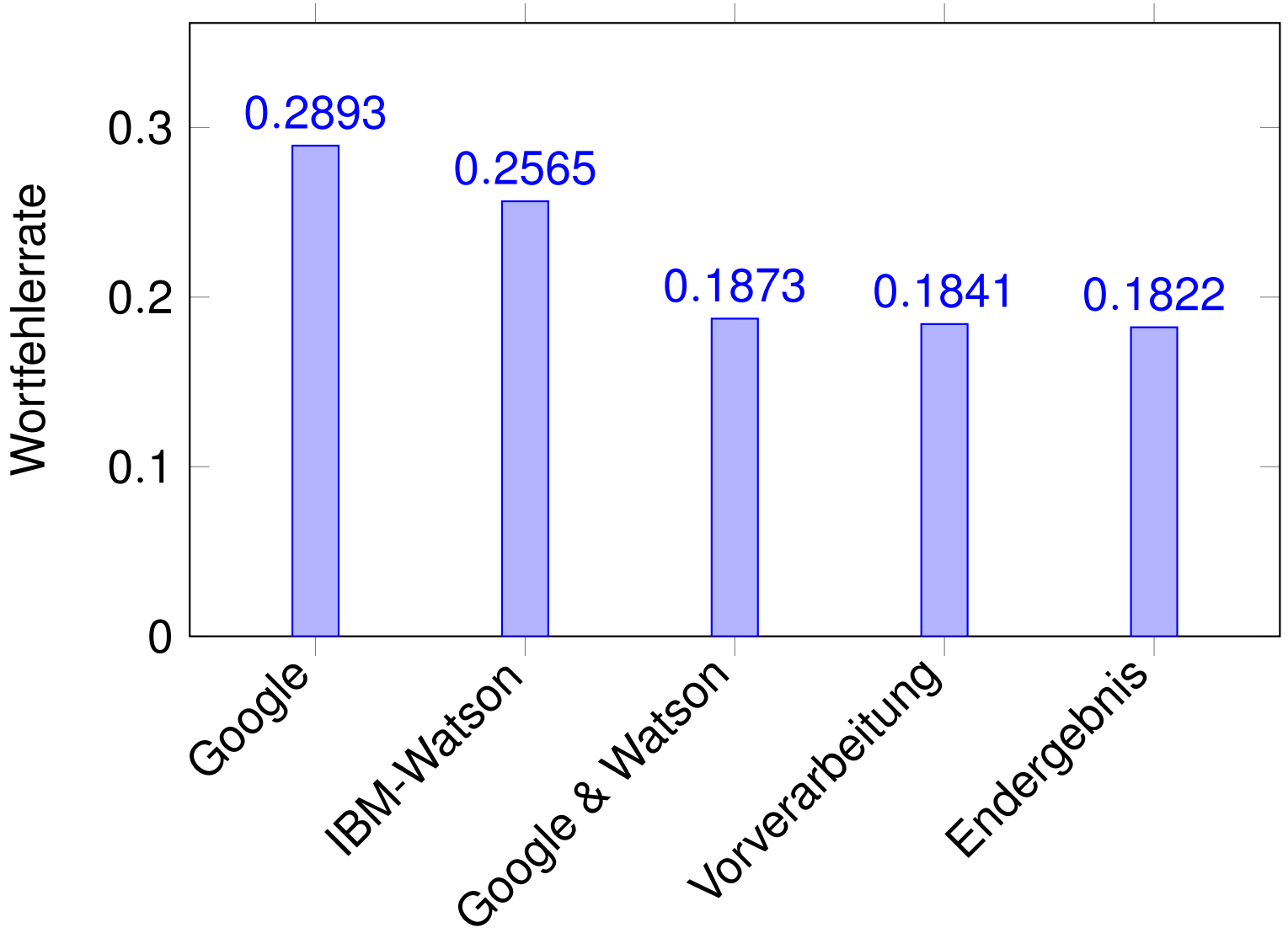
- Kombination mehrerer Ausgaben auch Spracherkennerübergreifend
- Kompakte Darstellung der Spracherkennerausgaben

- Eigenschaften
  - Gerichtet
  - Beschriftet
  - Zusammenhängend
  - Kreisfrei
  - Alle Pfade starten an einem Quellknoten.
  - Alle Pfade enden in einem Senkenknoten.
  - Alle Pfade gehen durch alle Knoten.

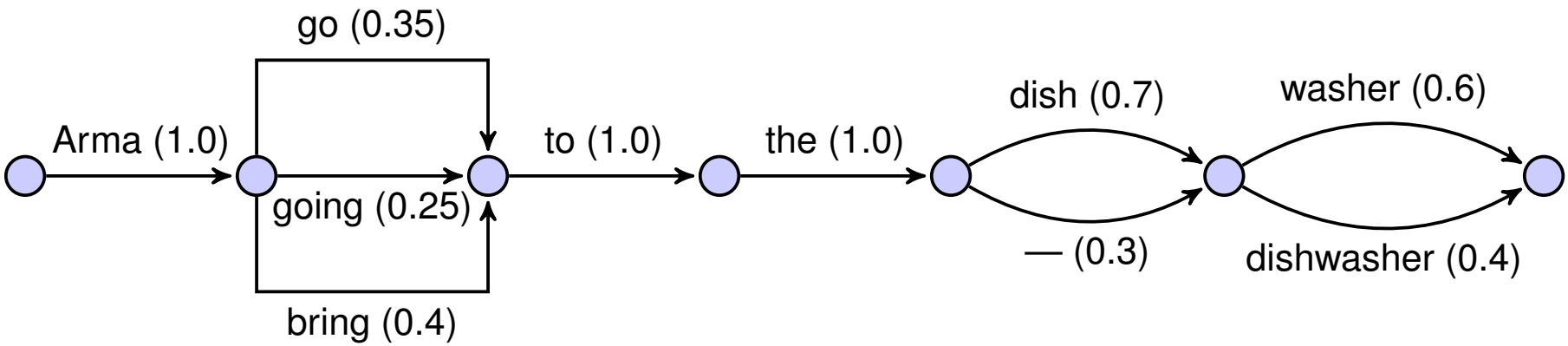
Lösungsansatz



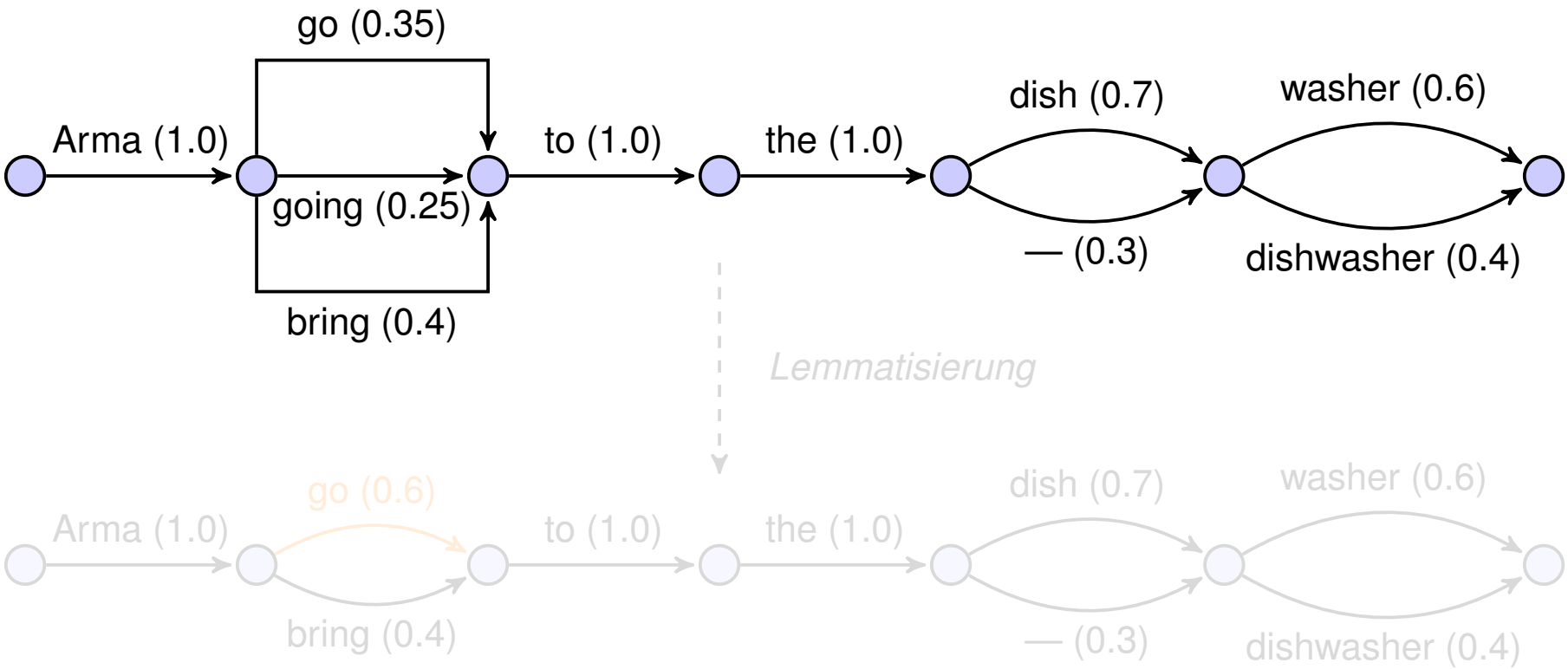
Lösungsansatz



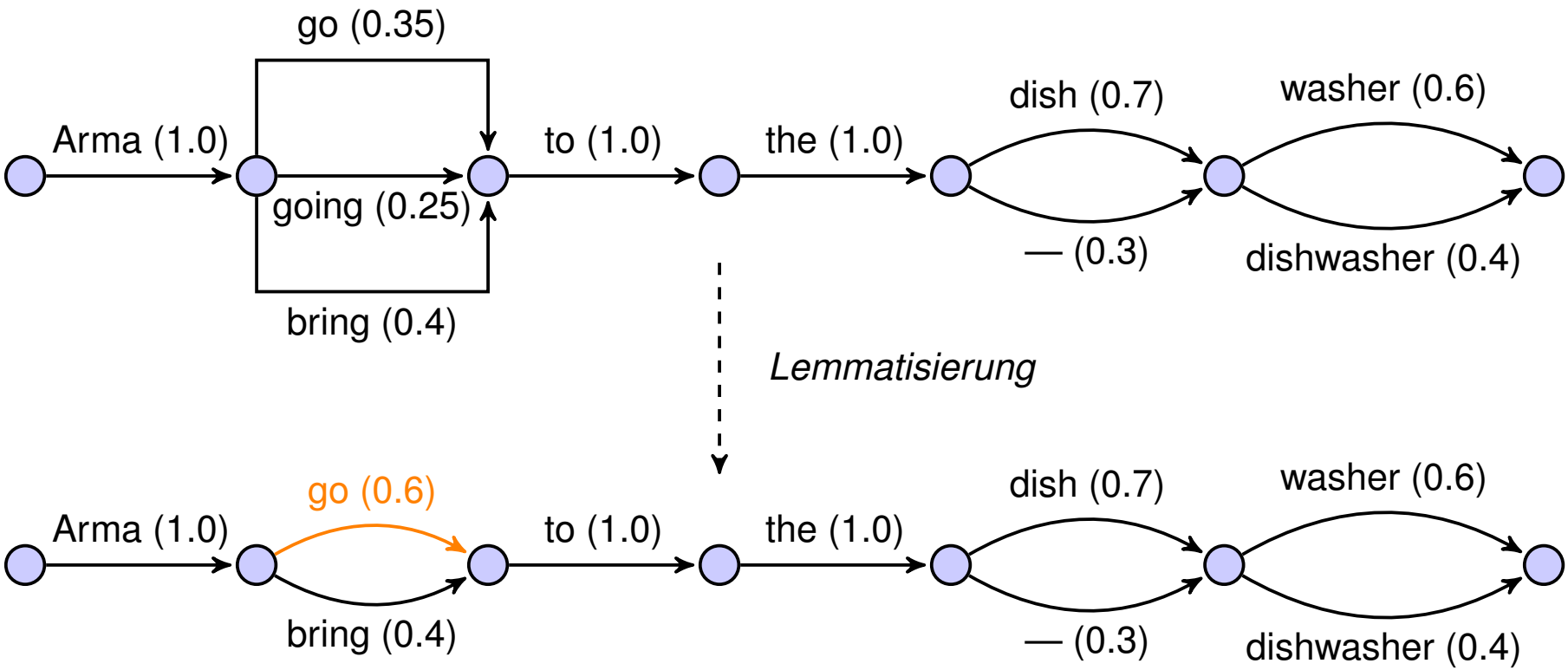
Ausgangsnetzwerk



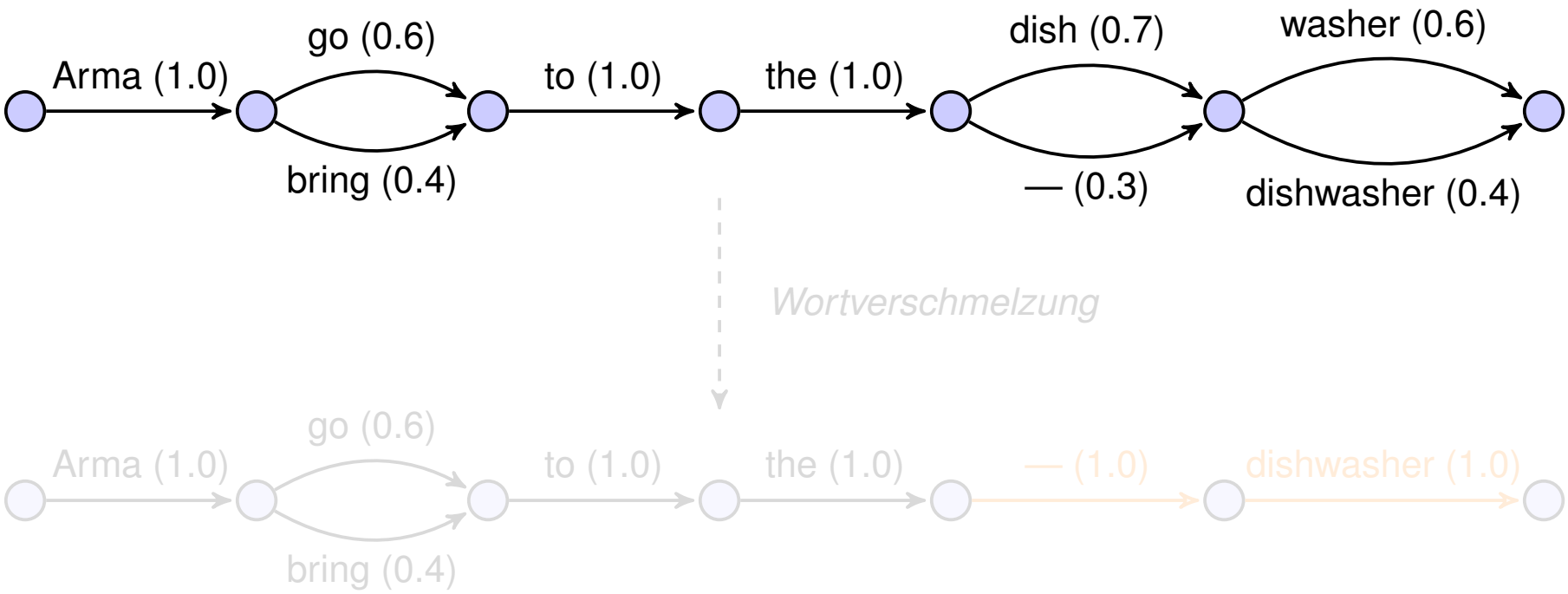
Vorverarbeitung – Lemmatisierung



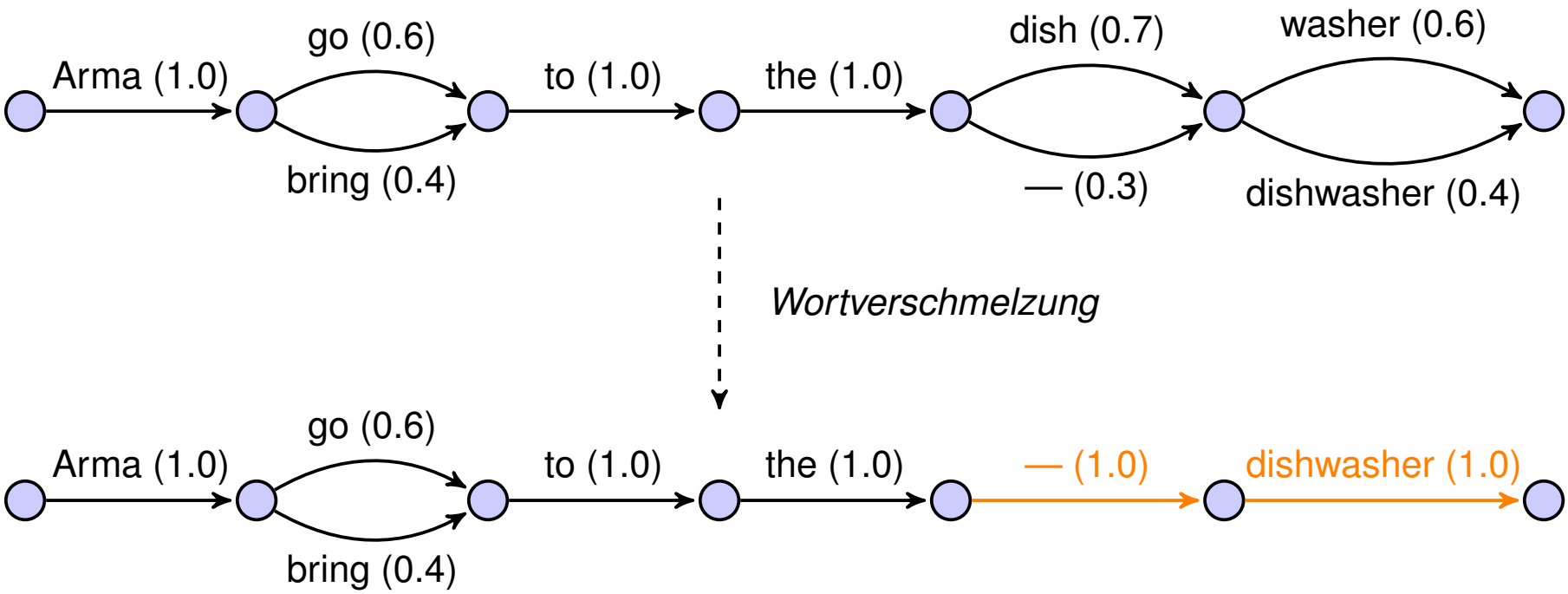
Vorverarbeitung – Lemmatisierung



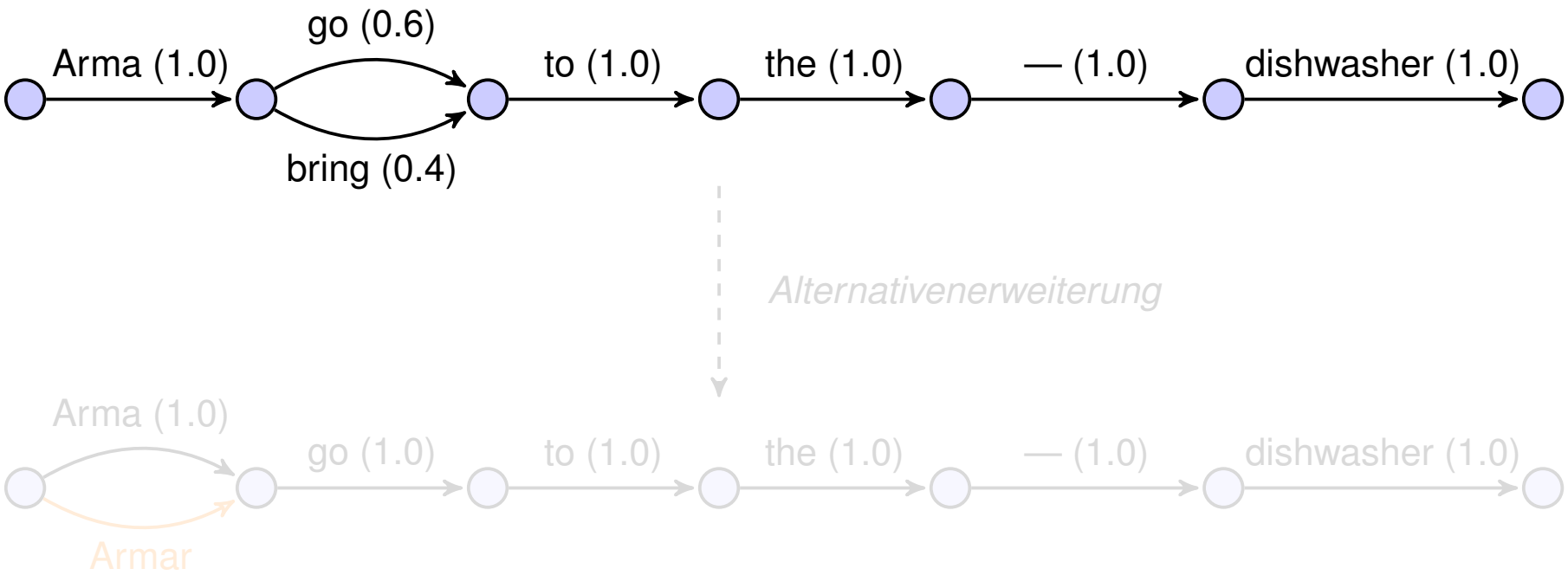
Vorverarbeitung – Wortverschmelzung



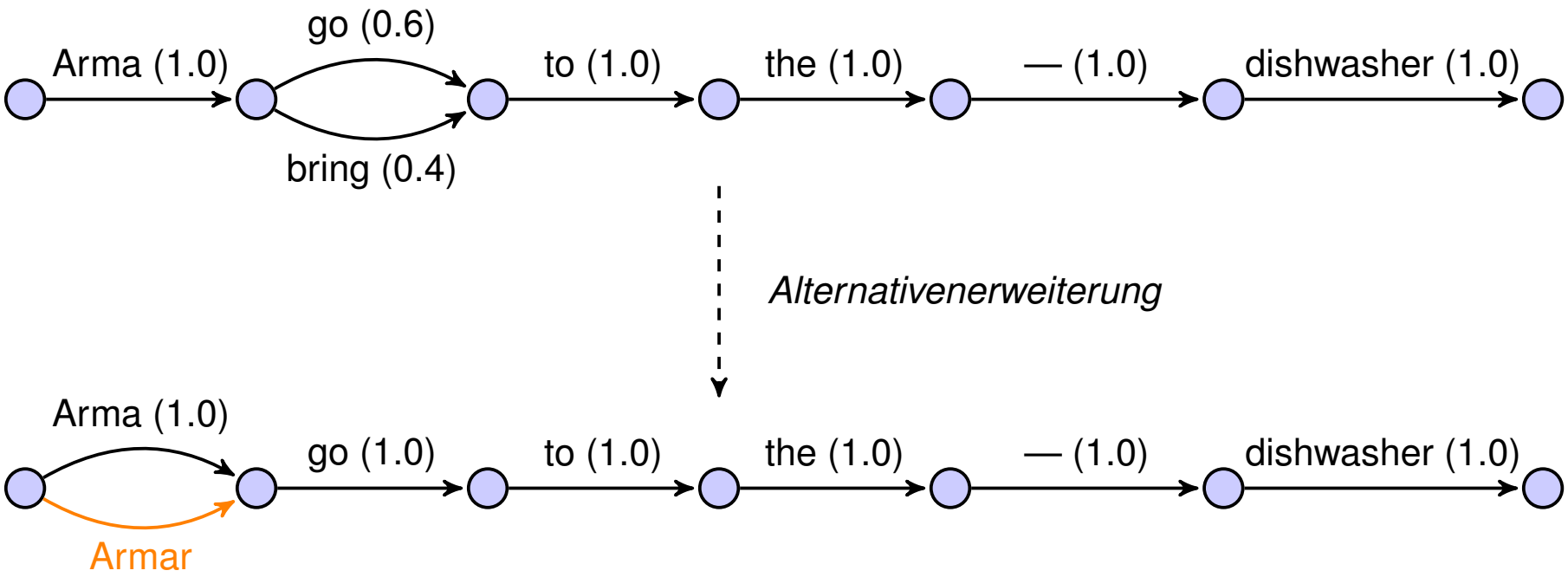
Vorverarbeitung – Wortverschmelzung



Vorverarbeitung – Alternativenerweiterung



Vorverarbeitung – Alternativenerweiterung



Phonetische Algorithmen

Phonetischer Algorithmus

- Versucht die Aussprache zu abstrahieren
- Double Metaphone [Phi00]
- “site” und “sight” $\rightarrow$ “ST” und “ST”.

Phonetische Algorithmen

Phonetischer Algorithmus

- Versucht die Aussprache zu abstrahieren
- Double Metaphone [Phi00]
- “site” und “sight” $\rightarrow$ “ST” und “ST”.

Phonetische Algorithmen

Phonetischer Algorithmus

- Versucht die Aussprache zu abstrahieren
- Double Metaphone [Phi00]
- “site” und “sight” $\rightarrow$ “ST” und “ST”.

Alternativenerweiterung

Beispiel

- Arma $\rightarrow$ ARM & ARM
- Armar $\rightarrow$ ARMR & ARMR
- Levensthein-Distanz 1 [Lev66]
- Jaro-Winkler-Distanz 94 [Win06]

Gegenbeispiele

- brews bruise $\rightarrow$ Levensthein-Distanz 3 — Jaro-Winkler-Distanz 0.2
- k s $\rightarrow$ Levensthein-Distanz 1 — Jaro-Winkler-Distanz 0.0

Alternativenerweiterung

Beispiel

- Arma $\rightarrow$ ARM & ARM
- Armar $\rightarrow$ ARMR & ARMR
- Levensthein-Distanz 1 [Lev66]
- Jaro-Winkler-Distanz 94 [Win06]

Gegenbeispiele

- brews bruise $\rightarrow$ Levensthein-Distanz 3 — Jaro-Winkler-Distanz 0.2
- k s $\rightarrow$ Levensthein-Distanz 1 — Jaro-Winkler-Distanz 0.0

Alternativenerweiterung

Beispiel

- Arma $\rightarrow$ ARM & ARM
- Armar $\rightarrow$ ARMR & ARMR
- Levensthein-Distanz 1 [Lev66]
- Jaro-Winkler-Distanz 94 [Win06]

Gegenbeispiele

- brews bruise $\rightarrow$ Levensthein-Distanz 3 — Jaro-Winkler-Distanz 0.2
- k s $\rightarrow$ Levensthein-Distanz 1 — Jaro-Winkler-Distanz 0.0

Bewertungsmodule

- Numerische Bewertung der Alternativen
- Pseudo-Sprachmodell

Bewertungsmodule

Verwirrungsnetzwerk

- Übergangswahrscheinlichkeit

Bewertungsmodule

Domäne

- Domänenontologie
- Roboteraktionen
- Namen
- Umfeld

Bewertungsmodule

WordNet

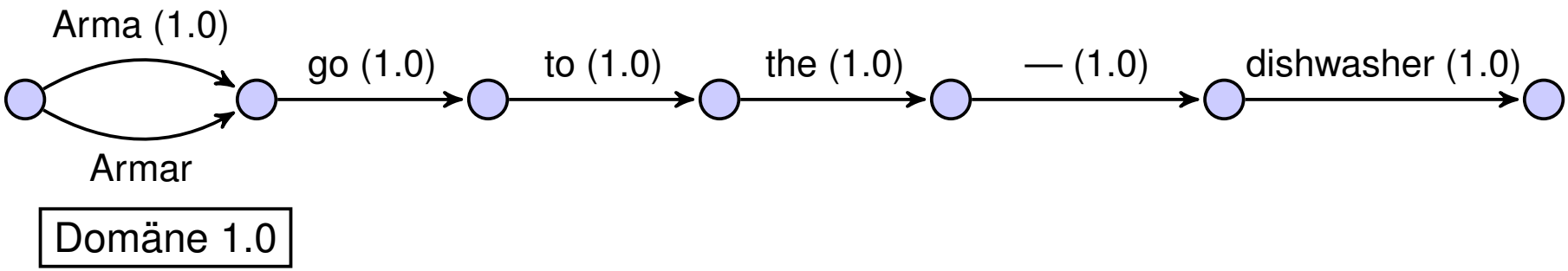
- Abstandsmaßen basierend auf WordNet
- Bewertung von Wortpaaren

Bewertungsmodule

Google

- Suchergebnissanzahl
- Bewertung von Wortpaaren

Ergebnisnetzwerk



Gewichte

- Gewicht pro Bewertungsmodul
- Kreuzvalidierung
- Differenzialevolution

Gewichte

- Gewicht pro Bewertungsmodul
- Kreuzvalidierung
- Differenzialevolution

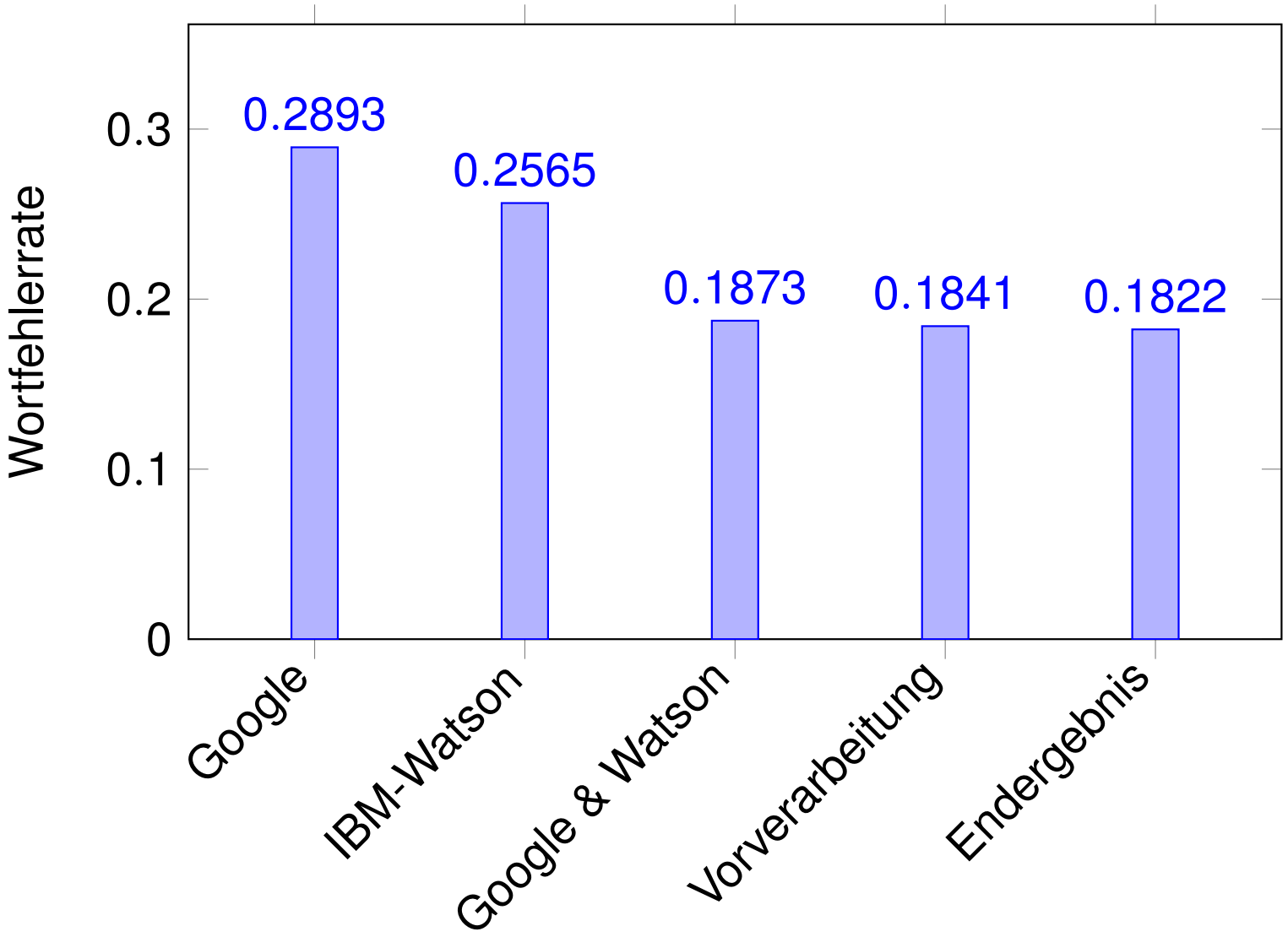
Gewichte

- Gewicht pro Bewertungsmodul
- Kreuzvalidierung
- Differenzialevolution

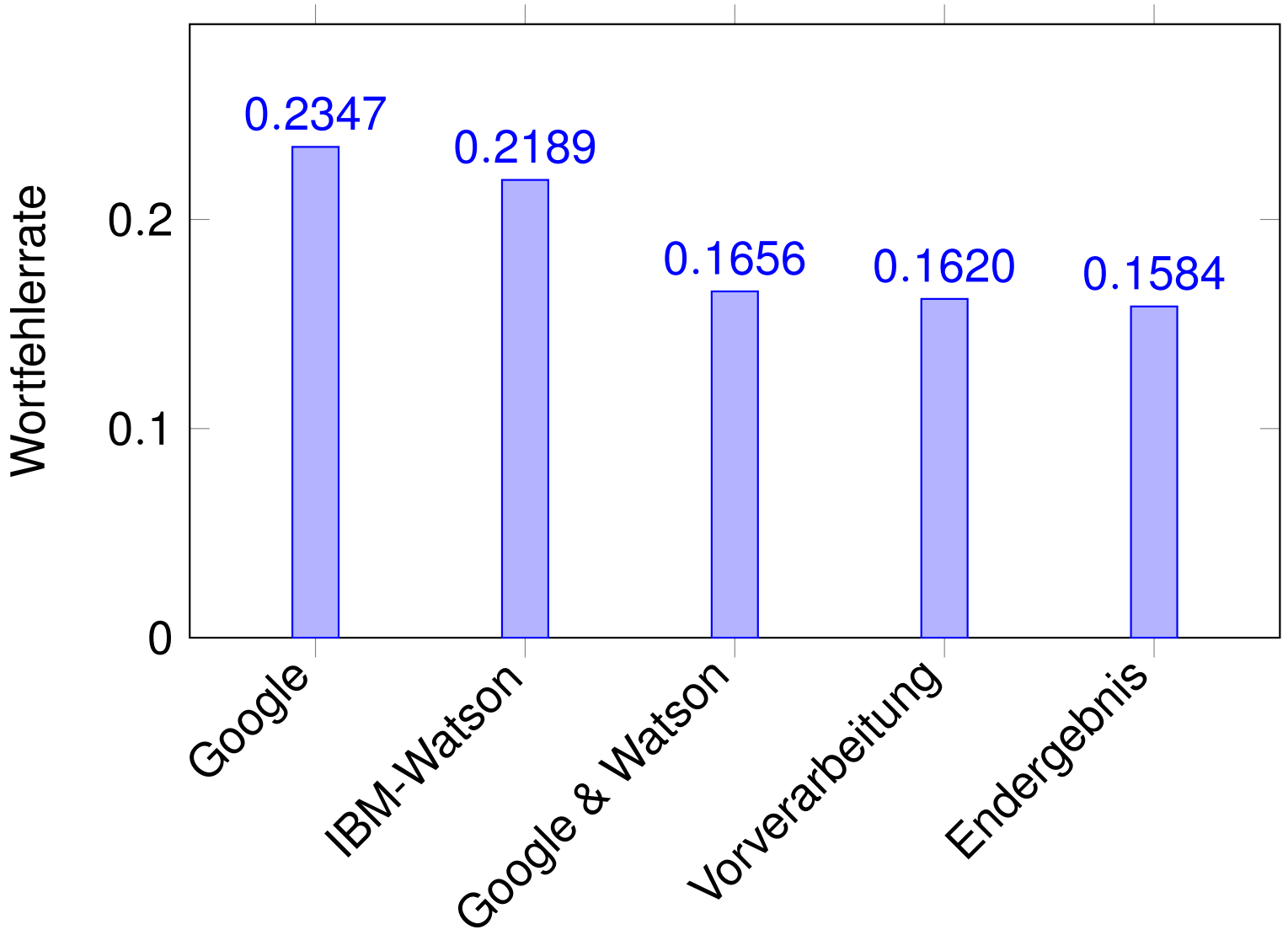
Korpus

Szenario	Kurzbeschreibung der Aufgabenstellung	#	Sprachk. ∅	Qualität	Quelle
1.	Armar soll Popcorn holen.	22	2.77	+	Günes [Gü15]
		14	2.14	-	Paskaran [Pas15]
2.	Armar soll einen Becher in die Spülmaschine stellen.	26	2.77	+	Günes [Gü15]
		14	2.14	-	Paskaran [Pas15]
3.	Armar soll Saft bringen.	23	2.77	+	Günes [Gü15]
		14	2.14	-	Paskaran [Pas15]
4.	Armar soll die Spülmaschine aus- und einräumen.	19	2.22	+	Steurer [Ste16]
5.	Armar soll einen Drink mixen.	19	2.22	+	Steurer [Ste16]

Evaluation - Kompletter Korpus



Evaluation - Szenario 1–3, RØDE NT1-A



Probleme

- Gebunden an die Spracherkenner
- Optimum? $\rightarrow$ Ranglernverfahren?

Ausblick

Unstetigkeiten

- Verzögerungslaute
  - ähm, mhh (eh, ehm, uhm)
- Korrekturen
  - Wiederholungen

Befehlsgrenzen

- Semantisch abgeschlossener Teil des Gesprochenen.
- „Bring mir den Saft und schließe die Tür.“

Ausblick

Unstetigkeiten

- Verzögerungslaute
  - ähm, mhh (eh, ehm, uhm)
- Korrekturen
  - Wiederholungen

Befehlsgrenzen

- Semantisch abgeschlossener Teil des Gesprochenen.
- „Bring mir den Saft und schließe die Tür.“

References I

Bangalore, Srinivas, German Bordel und Giuseppe Riccardi (2001). “Computing consensus translation from multiple machine translation systems”. In: IEEE, S. 351–354. ISBN: 0-7803-7343-X.

Fiscus, Jonathan G (1997). “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)”. In: *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, S. 347–354.

Günes, Zeynep (2015). “Aufbau eines Sprachkorpus zur Programmierung autonomer Roboter mittels natürlicher Sprache”. Bachelor’s Thesis. Karlsruher Institut für Technologie (KIT) – IPD Tichy. URL: `https://code.ipd.kit.edu/weigelt/parse/wikis/Theses/guenes_ba`.

Keller, Frank und Mirella Lapata (2003). “Using the web to obtain frequencies for unseen bigrams”. In: *Computational linguistics* 29.3, S. 459–484.

References II

- Levenshtein, Vladimir I (1966). “Binary codes capable of correcting deletions, insertions and reversals”. In: *Soviet physics doklady*. Bd. 10, S. 707.
- Mangu, Lidia, Eric Brill und Andreas Stolcke (2000). “Finding consensus in speech recognition: word error minimization and other applications of confusion networks”. In: *Computer Speech & Language* 14.4, S. 373–400. ISSN: 0885-2308.
- Matusov, Evgeny, Nicola Ueffing und Hermann Ney (2006). “Computing Consensus Translation for Multiple Machine Translation Systems Using Enhanced Hypothesis Alignment”. In:
- Nakov, Preslav und Marti Hearst (2005). “A study of using search engine page hits as a proxy for n-gram frequencies”. In: Citeseer.

References III

Paskaran, Dinesh (2015). “Evaluation unterschiedlicher Spracherkennungssysteme in der Domäne Humanoide Robotik”. Bachelor’s Thesis. Karlsruher Institut für Technologie (KIT) – IPD Tichy. URL: https://code.ipd.kit.edu/weigelt/parse/wikis/Theses/paskaran_ba.

Philips, Lawrence (2000). “The Double Metaphone Search Algorithm”. In: *C/C++ Users Journal* June 2000.

Princeton University (2010). *WordNet*. URL: http://wordnet.princeton.edu.

Steurer, Vanessa (2016). “Strukturerkennung von Bedingungen in gesprochener Sprache”. Bachelor’s Thesis. Karlsruher Institut für Technologie (KIT) – IPD Tichy. URL: https://code.ipd.kit.edu/weigelt/parse/wikis/Theses/steurer_ba.

Winkler, William E (2006). “Overview of record linkage and current research directions”. In: *Bureau of the Census*. Citeseer.

References IV

Yu, D. und L. Deng (2014). *Automatic Speech Recognition: A Deep Learning Approach*. Springer London. ISBN: 978-1-4471-5779-3. URL: https://books.google.de/books?id=rUBTBQAAQBAJ.

Bewertungskombination & Skalierung

- Durchschnitt
- Summe

$$b_i^\star(x) = (g_i * b_i(x))^{e_i} \tag{1}$$

mit: Skalierte Bewertung $b_i^\star(x)$ (2)

Alternative $x \in X$ Menge aller Alternativen (3)

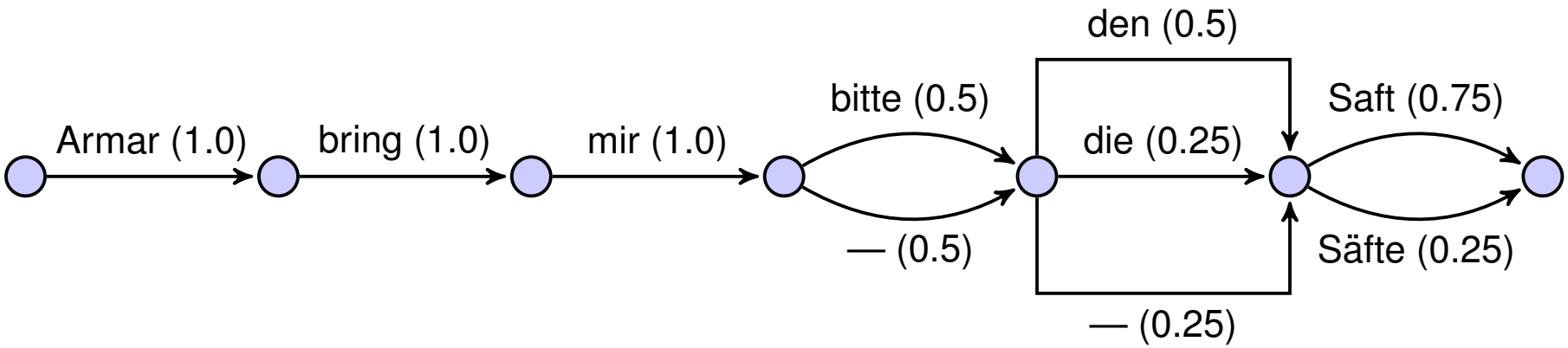
Bewertungsmodul $i \in I$ Menge aller Bewertungsmodule (4)

Beispielnetzwerk

Satz	Satzkonfidenz
Armar bring mir bitte den Saft	0.9
Armar bring mir den Saft	0.8
Armar bring mir bitte Saft	0.8
Armar bring mir die Säfte	0.7

Beispielnetzwerk

Satz	Satzkonfidenz
Armar bring mir bitte den Saft	0.9
Armar bring mir den Saft	0.8
Armar bring mir bitte Saft	0.8
Armar bring mir die Säfte	0.7



Evaluationsdaten I

Tabelle: Evaluationsergebnisse für den kompletten Korpus

Spracherkenner	WER	TERp	Bleu
1. Google	**0.2956 (0.1893)**	**0.2953 (0.1889)**	**0.6019 (0.6927)**
2. IBM Watson	0.2734	0.2715	0.6060
3. Google oder Watson	**0.2138**	**0.2132**	**0.6643**
4. Google CN	0.2965 (0.1904)	0.2962 (0.1900)	0.5991 (0.6895)
5. Watson CN	0.2790	0.2772	0.5917
6. Google oder Watson CN	0.2166	0.2160	0.6577
7. Google und Watson CN	0.2076	0.2072	0.6696
8. Revise CN	0.2047	0.2044	0.6746
9. Revise Alle	0.2029	0.2026	0.6785
10. Revise Ohne Lesk und Lch	**0.2035**	**0.2032**	**0.6774**
11. Revise Alle Sofia-ML	0.2172	0.2169	0.6553
12. Revise Sofia-ML	0.2154	0.2150	0.6537

Evaluationsdaten II

Tabelle: Evaluationsergebnisse für das bessere Mikrophon

Spracherkenner	WER	TERp	Bleu
1. Google	**0.2893 (0.1763)**	**0.2889 (0.1758)**	**0.6112 (0.7084)**
2. IBM Watson	0.2565	0.2549	0.6138
3. Google oder Watson	**0.1947**	**0.1943**	**0.6840**
4. Google CN	0.2873 (0.1740)	0.2869 (0.1735)	0.6127 (0.7101)
5. Watson CN	0.2647	0.2631	0.5949
6. Google oder Watson CN	0.1959	0.1955	0.6804
7. Google und Watson CN	0.1873	0.1873	0.6881
8. Revise CN	0.1841	0.1841	0.6949
9. Revise Alle	0.1822	0.1822	0.6967
10. Revise ohne Lch und Lesk	**0.1806**	**0.1806**	**0.6998**
11. Revise Alle Sofia-ML	0.2005	0.2005	0.6682
12. Revise Sofia-ML	0.2009	0.2009	0.6614

Evaluationsdaten III

Tabelle: Evaluationsergebnisse für das professionelle Mikrophon für Szenario 1–3

Spracherkenner	WER	TERp	Bleu
1. Google	**0.2347 (0.1715)**	**0.2347 (0.1715)**	**0.6614 (0.7161)**
2. IBM Watson	0.2189	0.2189	0.6498
3. Google oder Watson	**0.1807**	**0.1807**	**0.7012**
4. Google CN	0.2333 (0.1699)	0.2333 (0.1699)	0.6600 (0.7146)
5. Watson CN	0.2210	0.2210	0.6379
6. Google oder Watson CN	0.1785	0.1785	0.7011
7. Google und Watson CN	0.1656	0.1656	0.7079
8. Revise CN	0.1620	0.1620	0.7161
9. Revise Alle	0.1584	0.1584	0.7231
10. Revise ohne Lch und Lesk	**0.1569**	**0.1569**	**0.7261**
11. Revise Alle Sofia-ML	0.1605	0.1605	0.7186
12. Revise Sofia-ML	0.1641	0.1641	0.7098

Evaluationsdaten IV

Tabelle: Evaluationsergebnisse für Szenario 4 & 5 sowie Aufnahmen mit unbekanntem Mikrophon

Spracherkenner	WER	TERp	Bleu
1. Google	**0.3421 (0.2045)**	**0.3416 (0.2039)**	**0.5563 (0.6727)**
2. IBM Watson	0.3152	0.3118	0.5724
3. Google oder Watson	**0.2391**	**0.2380**	**0.6360**
4. Google CN	0.3449 (0.2079)	0.3444 (0.2072)	0.5525 (0.6680)
5. Watson CN	0.3234	0.3201	0.5563
6. Google oder Watson CN	0.2457	0.2446	0.6244
7. Google und Watson CN	0.2397	0.2391	0.6403
8. Revise CN	0.2375	0.2369	0.6429
9. Revise All	**0.2391**	**0.2386**	**0.6423**

Evaluationsdaten V

Tabelle: Evaluationsergebnisse mit Google Bewertungsmodul

Spracherkenner	WER	TERp	Bleu
1. Google	**0.3634 (0.2343)**	**0.3634 (0.2343)**	**0.5273 (0.6343)**
2. IBM Watson	0.3895	0.3837	0.5260
3. Google oder Watson	**0.3110**	**0.3052**	**0.5654**
4. Google CN	0.3750 (0.2483)	0.3750 (0.2483)	0.5110 (0.6146)
5. Watson CN	0.3953	0.3895	0.5206
6. Google oder Watson CN	0.3227	0.3169	0.5490
7. Google und Watson CN	0.3052	0.2994	0.5721
8. Revise CN	0.3081	0.2023	0.5671
10. Revise Alle	**0.3081**	**0.3052**	**0.5731**
10b. Revise Alle + Google	**0.3140**	**0.3081**	**0.5676**

Evaluationsdaten VI

Tabelle: Evaluationsergebnisse für den kompletten Korpus mit einem IBM-Watson Malus

Spracherkenner	WER	TERp	Bleu
Ganzer Korpus			
7. Google und Watson	0.2079	0.2076	0.6697
10. Revise Ohne Lesk und Lch	**0.2041**	**0.2038**	**0.6773**
Professionelles Mikrophon			
7. Google und Watson	0.1880	0.1880	0.6881
10. Revise Ohne Lesk und Lch	**0.1826**	**0.1826**	**0.6975**
Szenario 1–3 – Professionelles Mikrophon			
7. Google und Watson	0.1656	0.1656	0.7079
10. Revise Ohne Lesk und Lch	**0.1605**	**0.1605**	**0.7205**
Szenario 4–5 & unbekanntes Mikrophon			
7. Google und Watson	0.2402	0.2397	0.6405
10. Revise Ohne Lesk und Lch	**0.2424**	**0.2419**	**0.6356**

Ontologie I

- Thing
  - Method
    - bring
    - open
    - come
    - empty
    - fill
    - get
    - go
    - hand
    - open
    - pour
    - put
    - show
  - Object
    - cup

Ontologie II

- cupboard
- dishwasher
- door
- fridge
- juice
- microwave
- popcorn
- table
- television
- tv
- vodka

- Robot
  - Armar