WORTSINNDISAMBIGUIERUNG

Lars Mocka & Stefanie Schatohin-Edstein

Computerlinguistik I – WiSe 2021/2022

GLIEDERUNG

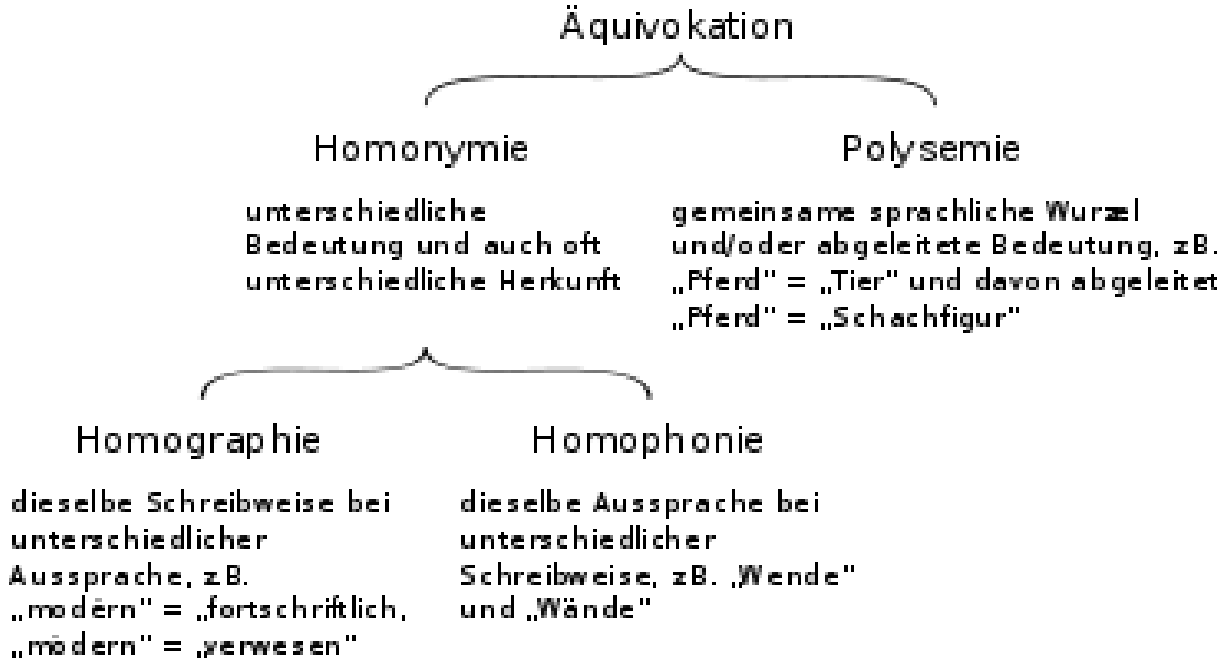
- Begriffserklärung
  - Beispiele
- Anwendungsbereiche
- Anfänge von WSD
- Ansätze und Methoden
  - Wörterbuch- und wissensbasierte Methoden
  - Überwachte Methoden
  - Semi-überwachte Methoden
  - Unüberwachte Methoden
- Aktueller Stand
- Beispiel: Lesk-Algorithmus

BEGRIFFSERKLÄRUNG

- **Ambiguität**: Mehrdeutigkeit
- **Disambiguierung**: Auflösen der sprachlichen Mehrdeutigkeit

  → jedoch nur mit Kontext möglich

- Identifizierung **des Sinnes eines Wortes**, wenn **mehrere** Bedeutungen vorliegen
- **Lexikalische** Mehrdeutigkeit: Vorhandensein von zwei oder mehr möglichen Bedeutungen in einem **einzelnen Wort**
- **Syntaktische** Mehrdeutigkeit: Vorhandensein von zwei oder mehr möglichen Bedeutungen innerhalb eines **einzelnen Satzes** oder einer Folge von Wörtern
- Bei Fällen von **Homonymie** geeignet



https://de.m.wikipedia.org/wiki/Datei:Homonymie.svg

BEISPIELE

Wortbeispiele:

- Bank: zum Sitzen vs. fürs Geld
- Absatz: im Text vs. am Schuh
- Rock: Kleidungsstück vs. Musikrichtung

Satzbeispiele:

- Der Jäger traf den Mann mit dem Gewehr.
- Ich bekam gestern einen Bienenstich.

ANWENDUNGSBEREICHE

- **Maschinelle Übersetzung**

  → change: Wandel oder Wechselgeld?

- **Information Extraction**

  → Informationen aus unstrukturierten Daten extrahieren

- **Information Retrieval**

- in anderen **computerbezogenen Schriften** weiterarbeiten

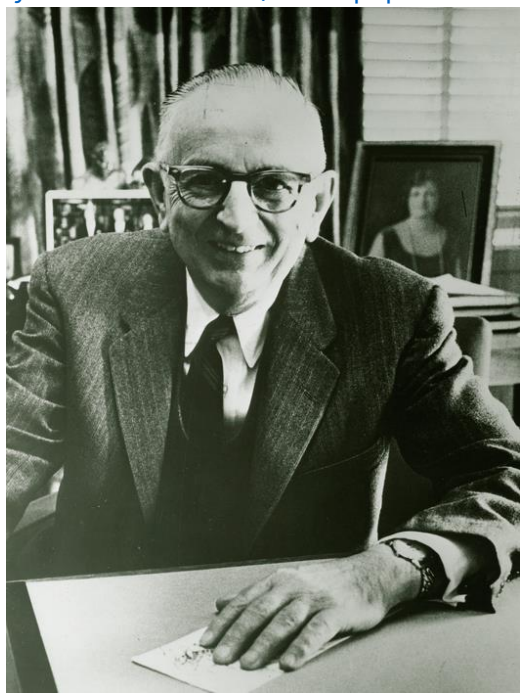
  → Verbesserung der Suchmaschinen

  → Inferenz

  → Kohärenz

https://www.historyofinformation.com/detail.php?id=2990

ANFÄNGE VON WSD



- In der maschinellen Übersetzung in den **1940er** Jahren als eigenständige Rechenaufgabe
  → eines der ältesten Probleme der Computerlinguistik
- Erstmals von **Warren Weaver** 1949 in seinem Memorandum über die Übersetzung in einen rechnerischen Kontext gesetzt.
- **1960 Bar-Hillel:** WSD kann nicht durch "elektronische Computer" gelöst werden, da im Allgemeinen das gesamte Weltwissen modelliert werden müsse.
- **1980er:** lexikalische Ressourcen wie das Oxford Advanced Learner's Dictionary of Current English (OALD) verfügbar
  → **Handcodierung** durch automatisch aus diesen Ressourcen extrahiertes Wissen ersetzt
  → **Begriffsklärung** immer noch wissensbasiert oder wörterbuchbasiert

ANSÄTZE UND METHODEN

1. Wörterbuch- und wissensbasierte Methoden
2. Überwachte Methoden
3. Semi-überwachte Methoden
4. Unüberwachte Methoden

WÖRTERBUCH- UND WISSENSBASIERTE METHODEN

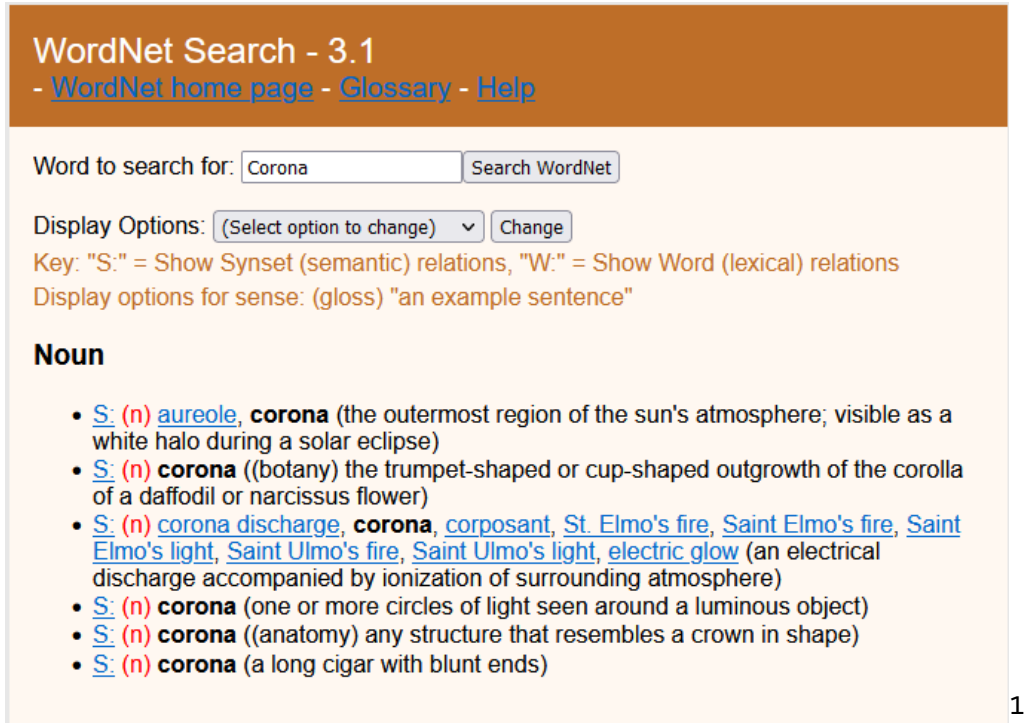
Englischer Begriff: Knowledgebased WSD / Unsupervised WSD

Stützen sich hauptsächlich auf **Wörterbücher**, **Thesauri, semantische Netze** und lexikalische **Wissensdatenbanken**, ohne auf Korpora zurückzugreifen.

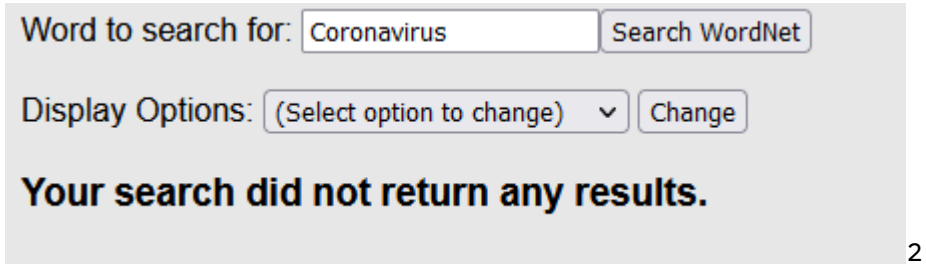
Voraussetzung: Es steht kein annotiertes Korpus zur Verfügung.

Problem: Wörterbücher sind statisch, Sprache ist aber immer im Wandel.

BEISPIEL FÜR (MÖGLICHERWEISE) VERALTETE WORTLISTEN



1



2

Quellen (Abgerufen am 10.01.2022):

1. http://wordnetweb.princeton.edu/perl/webwn?s=Corona&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=000000
2. http://wordnetweb.princeton.edu/perl/webwn?s=Coronavirus&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=000000

ÜBERWACHTE METHODEN

Englischer Begriff: Supervised WSD

Diese verwenden sinnvoll **kommentierte Korpora** zum Trainieren.

Voraussetzung: Es liegt bereits ein disambiguiertes Trainingskorpus vor. Wörter werden vorher klassifiziert!

Ziel: Die Maschine soll neue ambige Wörter erkennen und diese unter Berücksichtigung des Kontextes disambiguieren.

ÜBERWACHTE METHODEN

Umsetzungen:

- **Purely Data Driven**: arbeitet nur mit dem Korpus.
- **Supervised WSD Exploiting Glosses**: nutzt Textglossen, um den Kontext einzugrenzen
- **Supervised WSD Exploiting Relations**: nutzt semantische Netze. Hier werden die semantischen Beziehungen zwischen Begriffen dargestellt
- **Supervised WSD Exploiting Other Knowledge**: hier können zum Beispiel Bilder oder weitere Textquellen wie Wikipedia miteinbezogen werden.

SEMI-ÜBERWACHTE METHODEN

Englischer Begriff: Semi-Supervised WSD

Diese nutzen eine **sekundäre Wissensquelle** wie ein kleines annotiertes Korpus als Seed-Daten in einem Bootstrapping-Prozess oder ein wortorientiertes zweisprachiges Korpus.

UNÜBERWACHTE METHODEN

Englischer Begriff: Unsupervised WSD

Diese meiden (fast) vollständig externe Informationen und arbeiten direkt aus **rohen, nicht kommentierten Korpora**. Diese Verfahren sind auch unter dem Namen Wortsinn-Diskriminierung bekannt.

AKTUELLER STAND DER WSD

Wörterbuchbasierte Methoden werden seltener angewandt. Meistens eher in Semiüberwachten Verfahren.

Maschinelle Lernverfahren in der WSD nehmen zu.

Generell gibt es einen Trend, WSD nicht nur für die englische Sprache, sondern viele verschiedene Sprachen zu verwenden. Dabei geht es auch darum, eine multilinguale WSD zu schaffen (Beispiel: BabelNET).

LESK-ALGORITHMUS: ÜBERSICHT

Der Algorithmus wurde von Michael E. Lesk im Jahr 1986 entwickelt.

Es ist ein wissensbasierter Ansatz!

Es gibt nicht DEN einen Lesk-Algorithmus. Inzwischen gibt es verschiedene Variationen (Mal mit Wörterbuch, mal Mit semantischen Netz).

Probleme: Der Algorithmus ist ungenau.

LESK-ALGORITHMUS: BEISPIELANWENDUNG SCHRITT 1

- Zunächst suchen wir uns ein **Zielwort** aus, das wir disambiguieren möchten.
  - Wir nehmen das Wort **Maus**.
- Für **Maus** gibt es zwei Lesarten:
  - „**Maus**$_1$: kleines [graues] Nagetier mit spitzer Schnauze, das [als Schädling] in menschlichen Behausungen, auf Feldern u. in Wäldern lebt.
  - **Maus**$_2$: meist auf Rollen gleitendes, über ein Kabel mit einem PC verbundenes Gerät, das auf dem Tisch hin u. her bewegt wird, um den Cursor od. ein anderes Markierungssymbol auf dem Monitor des Computers zu steuern.“ (Quelle auf der Quellenfolie).

LESK-ALGORITHMUS: BEISPIELANWENDUNG SCHRITT 2

- Beispielsätze mit dem Zielwort **Maus**:
  a. Ein *Klick* mit der **Maus$_2$** und der *Computer* zaubert ein Video auf den *Monitor*.
  b. Mäuse$_1$ begeistern weltweit als Comic- und *Zeichentrickfiguren* ein riesiges Publikum, gleichzeitig werden sie als *Schädlinge gejagt*.
  c. Auch hier ersetzt das *Touchpad* die **Maus$_2$**.
  d. Da war eine **Maus$_1$**, die ein Kabel *angeknabbert* hat.
  e. Die **Maus$_1$** hat bei der Schulkameradin inzwischen ein neues Zuhause gefunden.

- Die *kursiven* Wörter geben menschlichen Leser*innen schon erste Hinweise auf die gesuchte Bedeutung des Zielwortes.

LESK-ALGORITHMUS: BEISPIELANWENDUNG

- Die kursiven Wörter werden auch **Kotext** bezeichnet.
  - Kotext = Nur die Wörter in unmittelbarer Nähe des Zielwortes. D.h. Maximal zwei Wörter sind Kotext. Steht das Zielwort am Satzanfang/-Ende gibt es nur ein Wort, das im Kotext steht
- Der gesamte Satz ist der **Kontext**.

LESK-ALGORITHMUS: BEISPIELANWENDUNG SCHRITT 3

Nun werden alle **Kontextwörter** für jeden Satz aufgelistet. Dazu zählen auch die kursiven Hinweiswörter und die **Kotextwörter**!

Funktionswörter (wie z.B. *und*) werden nicht aufgelistet, da sie unspezifisch sind. Sie werden als **Stoppwörter** behandelt.

Für unsere Beispielsätze listen wir folgende Kontextwörter auf:

a. *Computer, Klick, Monitor, Video, zaubern*
b. *begeistern, Comicfigur, jagen, Publikum, Schädling, weltweit, Zeichentrickfigur*
c. *ersetzen, Touchpad*
d. *anknabbern, Kabel*
e. *finden, Schulkameradin, Zuhause*

LESK-ALGORITHMUS: BEISPIELANWENDUNG SCHRITT 4

Jetzt geht es darum, dass wir in den Kontextwörtern gute **Indikatorwörter** für die möglichen Lesarten finden.

Dazu betrachten wir uns erneut die Lesarten des Zielwortes. Wir lemmatisieren die Definitionen und entfernen Stoppwörter:

1. **Maus$_1$**: *Behausung, Feld, grau, klein, leben, menschlich, Nagetier, Schädling, Schnautze, spitz, Wald*
2. **Maus$_2$**: *bewegen, Computer, Cursor, Gerät, gleiten, Kabel, Markierungsymbol, Monitor, PC, Rolle, steuern, Tisch, verbinden.*

LESK-ALGORITHMUS: BEISPIELANWENDUNG SCHRITT 4

Wir gleichen nun die Kontextwörter aus den Beispielsätzen mit den Wörtern aus den Definitionen ab:

a. *Computer$_2$, Klick, Monitor$_2$, Video, zaubern*
b. *begeistern, Comicfigur, jagen, Publikum, Schädling$_1$, weltweit, Zeichentrickfigur*
c. *ersetzen, Touchpad*
d. *anknabbern, Kabel$_2$*
e. *finden, Schulkameradin, Zuhause*

Die unterstrichenen Wörter haben einen Index zugewiesen bekommen, je nachdem mit welcher Lesart sie übereinstimmen.

LESK-ALGORITHMUS: BEISPIELANWENDUNG

Bei Satz d findet sich eine falsche Zuordnung: *anknabbern, Kabel$_2$*

Kabel wurde der Lesart 2 zugeordnet.

Zur Erinnerung Lesart 2:

- „**Maus$_2$**: meist auf Rollen gleitendes, über ein Kabel mit einem PC verbundenes Gerät, das auf dem Tisch hin u. her bewegt wird, um den Cursor od. ein anderes Markierungssymbol auf dem Monitor des Computers zu steuern.“

Und nochmal der Satz:

- „Da war eine **Maus$_1$**, die ein Kabel *angeknabbert* hat“.

Das Wort **Kabel** führt den Algorithmus hier in die Irre!

LESK-ALGORITHMUS: BEISPIELANWENDUNG

Problem: Der Lesk-Algorithmus entscheidet danach, wie viele Übereinstimmungen es zwischen Definitionen und Kontext gibt. Es kann vorkommen, dass eine Definition mehr Indikatorwörter als die andere hat. Oder dass ein Satz mehr Kontextwörter enthält als ein anderer.

Daher hat sich der Algorithmus sich in unserem Beispiel für die falsche Lesart von **Kabel** entschieden.

LESK-ALGORITHMUS: BEISPIELANWENDUNG SCHRITT 5

Lesk löst das Problem, indem er auch die Kontextwörter im Wörterbuch nachschaut.

Im Beispielsatz c (Auch hier ersetzt das *Touchpad* die **Maus$_2$**.)

Für **ersetzen** finden wir zwei Lesarten:

1. **ersetzen$_1$**: *für jmdn./etw. Ersatz schaffen; jmdn./etw. an die Stelle von jmdn./etw. setzen; für jmdn./etw. ein Ersatz sein; an die Stelle von jmdn./etw treten*
2. **ersetzen$_2$**: *erstatten, wiedergeben, für etw. Ersatz leisten*

Für **Touchpad** finden wir folgende Lesart:

1. **Touchpad**: *auf Fingerdruck reagierendes, im Computer integriertes Zeigegerät zur Steuerung des Cursors anstelle einer Maus*

LESK-ALGORITHMUS: BEISPIELANWENDUNG

Wir wählen **ersetzen$_1$**, weil es mehr Überschneidungsmöglichkeiten gibt.

Wir filtern nun die relevanten Wörter aus den Definition von **ersetzen$_1$** und **Touchpad** und kombinieren sie mit der Kontextliste für den Satz.

So ergibt sich folgender **erweiterter Kontext**:

- *Computer$_2$, Cursor$_2$, Ersatz, ersetzen, Fingerdruck, integriert, Maus, reagieren, schaffen, setzen, Stelle, Steuerung, Touchpad, treten, Zeigegerät*

Die unterstrichenen Wörter stellen Überschneidungen mit **Maus$_2$** dar.

LESK-ALGORITHMUS: BEISPIELANWENDUNG SCHRITT 6

Nun müssen wir die Indikatorwörter aus den Maus-Definitionen mit dem erweiterten Kontext abgleichen.

Da es auch Zufallstreffer geben kann (z.B. *Kabel*), verwenden wir den **Dice-Koeffizienten**.

$$\textbf{Dice} = \frac{2 \cdot |\text{Indikatorwörter} \cap \text{erweiterte Kontextwörter}|}{|\text{Indikatorwörter}| + |\text{erweiterte Kontextwörter}|}$$

Carstensen et al., S.385

LESK-ALGORITHMUS: BEISPIELANWENDUNG DICE FÜR $MAUS_1$

Der Dice-Koeffizient für **$Maus_1$**:

$$\mathrm{Dice}(1) = \frac{2 * 0}{11 + 15} = \frac{0}{26} = 0$$

Da es 0 Überschneidungen zwischen den Indikatorwörtern aus **$Maus_1$** und dem erweiterten Kontext des Satzes gibt, tragen wir eine 0 ein.

Insgesamt gibt es 11 Indikatorwörter für **$Maus_1$** und 15 Wörter im erweiterten Kontext.

Der Dice-Koeffizient für **$Maus_1$** beträgt 0.

LESK-ALGORITHMUS: BEISPIELANWENDUNG DICE FÜR $MAUS_2$

Der Dice-Koeffizient für **$Maus_2$**:

$$\text{Dice}(2) = \frac{2 \cdot 2}{13 + 15} = \frac{4}{28} = \frac{1}{7}$$ Carstensen et al., S.385

Da es 2 Überschneidungen zwischen den Indikatorwörtern aus **$Maus_2$** und dem erweiterten Kontext des Satzes gibt, tragen wir eine 2 ein.

Insgesamt gibt es 13 Indikatorwörter für **$Maus_1$** und 15 Wörter im erweiterten Kontext.

Der Dice-Koeffizient für **$Maus_2$** beträgt $\frac{1}{7}$.

Da Dice(2) > Dice(1) ist, wählen wir **$Maus_2$** als Lesart.

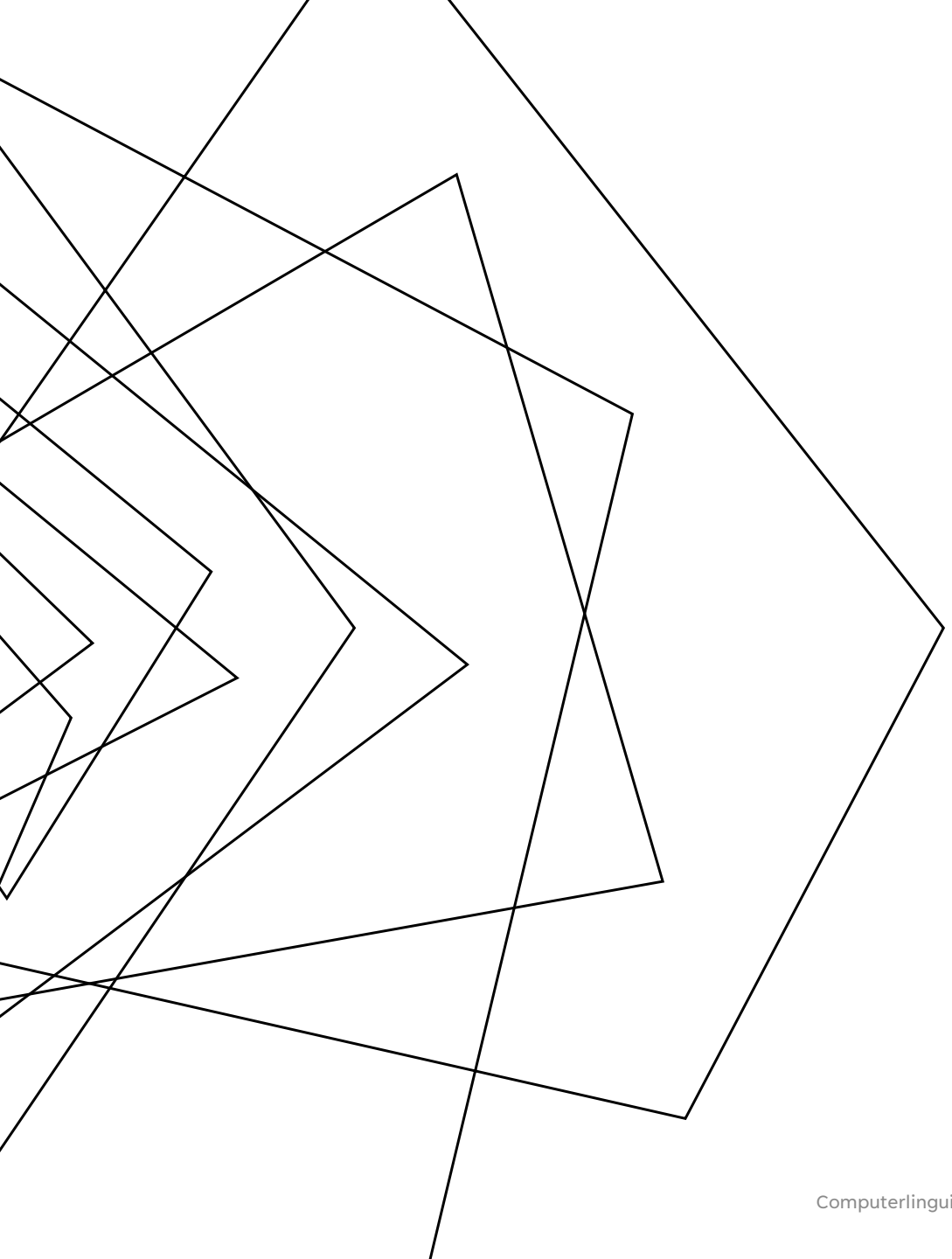
OPTIMIERUNG DES LESK-ALGORITHMUS

Der Lesk-Algorithmus wäre effizienter, wenn …

- …es in den Wörterbüchern auch immer schon Beispielsätze gäbe
- … in den Wörterbüchern schon die Häufigkeit für Indikator- und Kontextwörter angegeben würde (Diese Zahlen sollten dann aus Korpora stammen)
- … er mit Wortnetzen verwendet wird.

QUELLENVERZEICHNIS

- https://www.ibm.com/docs/en/db2/9.7?topic=studio-information-extraction
- https://de.wikipedia.org/wiki/Disambiguierung
- https://en.wikipedia.org/wiki/Word-sense_disambiguation
- http://docplayer.org/220814196-Word-sense-disambiguation-ueberblick.html
- http://docplayer.org/50085962-Wortbedeutungsdisambiguierung.html
- Die Übersicht über die verschiedenen WSD-Ansätze stammt aus Recent Trends in Word Sense Disambiguation: A Survey von Bevilacqua et al. (2021, Universität Helsinki)
- Das Beispiel für den Lesk-Algorithmus stammt aus Computerlinguistik und Sprachtechnologie – Eine Einführung, hrsg. von Carstensen et al. (3. Auflage, 2010, Spektrum Akademischer Verlag Heidelberg), S. 382-385
- WordNet: http://wordnetweb.princeton.edu/perl/webwn



VIELEN DANK FÜR EURE AUFMERKSAMKEIT!