Psycholinguistische Perspektiven auf Ambiguität

Christian Wurm (Düsseldorf)

Düsseldorf, 2023

Wir haben gesehen: es gibt die (soweit ich weiß universelle) Korrelation

$$\text{Häufigkeit von } w \sim \text{Anzahl der Lesarten von } w$$

Das lässt einen wichtigen Schluss zu:

Ambiguität

Ambiguität ist kein zufälliges Phänomen natürlicher Sprachen, sondern ein wichtiges Merkmal.
Natürliche Sprachen versuchen nicht, Ambiguität zu vermeiden, sondern (bis zu einem gewissen Grad) sie **suchen** Ambiguität.

Wie machen wir uns hier einen Reim drauf?

Warum Ambiguität?

Das ist natürlich eine psycho-linguistische Fragestellung; dementsprechend gibt es mehrere Antworten. Wir schauen uns drei Arbeiten zum Thema an:

1. Cancho, Sole, nach Zipf: Speaker and Hearer Pressure
2. (Piantadosi et al.: Communicative Function of Ambiguity)
3. ((Frost-Arnold, Beebe: Confused Terms in Ordinary Language))

Die Fakten

Erstmal ein paar Fakten zu Ambiguität:

- Wir gehen davon aus, dass wir uns verstehen (also Kommunikation ist erfolgreich)
- Es gibt viel Ambiguität in natürlicher Sprache, aber nicht übertrieben viel: sie erlaubt es, normalerweise im Kontext uns zu verstehen.
- Es gibt zwei widerstreitende Kräfte, die natürliche Sprache formen, nicht ausschließlich, aber maßgeblich:
  1. Ökonomie: wir sind einfach faul, und wollen beim Sprechen möglichst wenig Arbeit
  2. Deutlichkeit/Redundanz: Redundanz bedeutet, dass das Signal auch bei (leichten) Störungen verständlich bleibt.

Ökonomie und Redundanz sind einerseits entgegengesetzt, andererseits beide gleich wichtig, um effiziente Kommunikation zu gewährleisten.

FSZ: Der Hörer und der Sprecher

FSZ unterscheiden **Ökonomie nach Perspektive**:

- Die Ökonomie des Sprechers: der Sprecher will möglichst wenig **sagen**.
- Die Ökonomie des Hörers: der Hörer will möglichst wenig **denken**.

Das FSZ-Szenario

Nimm an, S will einen Sachverhalt P kommunizieren, und H will ihn wissen. (Wir kommen dazu warum das problematisch ist).

- S will P möglichst kompakt machen (mit wenig Worten sagen)
- H will, dass P mit möglichst unambig kommuniziert wird (sonst muss er überlegen)

Hörer, Sprecher, Ambiguität

Es liegt auf der Hand wie das mit Ambiguität zusammenhängt:

Ceteris paribus

Je ambiger eine (interpretierte) Sprache ist, desto kürzer wird im Mittel die Äußerung sein, die nötig ist, um eine Bedeutung auszudrücken.

Warum? Je ambiger, desto weniger Sätze brauche ich für jede endliche Menge von Bedeutungen, also sind die Sätze im Mittel kürzer

Hörer, Sprecher, Ambiguität

Also:

- Als Sprecher möchte man eine möglichst ambige Sprache
- Als Hörer möchte man eine möglichst unambige Sprache

Hieraus lassen sich leicht zwei Extremfälle ableiten:

- L_A, die maximal ambige Sprache, mit einem Wort a, dass *alle* Bedeutungen hat
- L_U, die unambige Sprache

Natürliche Sprachen als Mittelweg

FSZ sagen nun, kurz gesagt,

Die FSZ Hypothese

Natürliche Sprachen liegen irgendwo zwischen L_A und L_U – und zwar so, dass Sie einen optimalen Kompromiss zwischen S und H bilden.

Wie findet man Evidenz für so was? FSZ machen ein Experiment in einer Spielzeug-Sprache.

Die Spielzeug-Sprache

Die Zutaten:

- Wir nehmen eine (endliche) Menge $\{a_1, ..., a_n\} = A$ von Ausdrücken.
- Wir nehmen eine (endliche) Menge $\{r_1, ..., r_m\} = R$ von Referenten.
- Wir nehmen eine Bedeutungsmatrix $\mathbf{B} \in \{0, 1\}^{n \times m}$

Und fertig:

$$b_{ij} = 1 \tag{1}$$

bedeutet: a_i "bedeutet" (u.a.) r_j

$$b_{ij} = 0 \tag{2}$$

bedeutet: wir haben hier keine Referenz.

Die Bedeutungsmatrix

Die Frage ist: wie soll die Bedeutungsmatrix aussehen? Es gibt nur eine Beschränkung:

1. Für alle $j : 1 \leq j \leq m$ gibt es ein $i : 1 \leq i \leq n$ so dass $b_{ij} = 1$ (jede Referenz wird bezeichnet)
2. Aber: nicht jeder Ausdruck muss eine Referenz haben!

Etwas graphischer: Keine Spalte der Matrix darf der 0-Vektor sein.

Was den Rest angeht: wir erlauben alles. Aber: wir wollen optimieren!

Das Optimierungsproblem

Für ein Optimierungsproblem brauchen wir erstmal Kosten (die optimiert werden). Wir unterscheiden:

1. Die **Sprecherkosten** KS
2. Die **Hörerkosten** KH

Am Ende möchten wir optimieren

$$\lambda KS + (1 - \lambda)KH \tag{3}$$

für $\lambda \in (0, 1)$, genauer gesagt: $\lambda = 0.5$ (aber die Autoren sind da etwas laxer)

Die Sprecherkosten

Die Sprecherkosten sind definiert als eine **Entropie**. Wir brauchen also Wahrscheinlichkeiten.

Die Wahrscheinlichkeiten sind definiert von den Referenten her, und wir nehmen an

$$p(r_j) = \frac{1}{m} \tag{4}$$

Also alle Referenten sind gleich wahrscheinlich. Wir tun das in Abwesenheit von Gründen, warum es anders sein sollte.

Das Problem der Synoyme

Nehmen wir an, es gäbe keine Synonyme. Dann hätten wir ganz flott die Wahrscheinlichkeit über A:

$$p(a_i) = \sum_j b_{ij} p(r_j) \tag{5}$$

Also die Summe der Wahrscheinlichkeiten der Referenten, die er bezeichnet.
Aber: a_i und $a_{i'}$ können beide denselben Referenten bezeichnen! Also wird das etwas komplizierter:

$$p(a_i|r_j) = \frac{1}{\sum_k b_{kj}} b_{ij} \tag{6}$$

(die Wahrscheinlichkeit, dass a_i für r_j verwendet wird, ist 1/Anzahl der Worte, die es bezeichnen)

Das Problem der Synomyme

Wir bekommen dann:

$$p(a_i, r_j) = p(a_i|r_j)p(r_j) = \frac{1}{\sum_k b_{kj}} b_{ij} p(r_j) \tag{7}$$

(Produktregel) und

$$p(a_i) = \sum_j p(a_i, r_j) \tag{8}$$

(Summenregel)
Sprich: die Wahrscheinlichkeit eines Ausdrucks ist die Summe der Wahrscheinlichkeiten, dass er in einem gewissen Sinn verwendet wird.

Kurze Zusammenfassung

1. alle r_j sind gleich wahrscheinlich (problematisch!)
2. gegeben r_j ist jeder Ausdruck, der r_j denotiert, gleich wahrscheinlich
3. $p(a_i, r_j)$ folgt dann aus der Produktregel
4. $p(a_i)$ folgt aus Summenregel/Marginalisierung

Wir sehen also: für die Wahrscheinlichkeiten muss man ein paar Annahmen machen, aber nicht allzu viele!

Die Sprecherkosten *KS*

Sei p_A die Wahrscheinlichkeitsverteilung über A

$$p_A : A \rightarrow [0, 1] \tag{9}$$

Wir definieren nun ganz einfach KS mit der Entropie:

$$KS = H(p_A) = -\sum_i p(a_i) log\ p(a_i) \tag{10}$$

Hier die beiden Extremfälle:

- Nimm an, a_1 bezeichnet $r_1, ..., r_m$, alle anderen $a_i : i > 1$ bezeichnen *nichts*. Dann ist $p(a_1) = 1$, und $KS = H(p_a) = 0$
- Nimm an, $m = n$ und **B** ist eine Einheitsmatrix (1-1 Zuordnung). Dann ist

  $$KS = log(1/n) \tag{11}$$

  was in diesem Fall *maximal* ist (uniforme Verteilung!)

Die Sprecherkosten *KS*

Hier die beiden Extremfälle:

- Nimm an, a_1 bezeichnet $r_1, ..., r_m$, alle anderen $a_i : i > 1$ bezeichnen *nichts*. Dann ist $p(a_1) = 1$, und $KS = H(p_a) = 0$
- Nimm an, $m = n$ und **B** ist eine Einheitsmatrix (1-1 Zuordnung). Dann ist

$$KS = \sum (1/n) log(1/n) \tag{12}$$

  was in diesem Fall *maximal* ist (uniforme Verteilung!)

Wir haben also eine untere und eine obere Schranke für die Kosten.

Unsere Matrix sollte irgendwo dazwischen sein!

Hörerkosten *KH*

Wir haben p_A die Verteilung über A genannt. Dementsprechend wird nun

$$p_R : R \rightarrow [0, 1]$$

die Verteilung über R, d.h.

$$p_R(r) = p(r) = \frac{1}{|R|} \tag{13}$$

Wir können dementsprechend definieren:

$$H(p_R) = \sum_j p_R(r_j) \cdot log\ p_R(r_j) \tag{14}$$

Das ist aber nicht was wir wollen: die Anstrengung des Hörers ergibt sich nicht aus der Entropie von p_R – er bekommt ja ein Signal!

Hörerkosten

Also: **bedingte Entropie**:

$$H(p_R|a_i) = \sum_j p_R(r_j|a_i) \cdot log\ p_R(r_j|a_i) \tag{15}$$

Und nun abstrahieren wir vom konkreten a_i:

$$H(p_R|A) = \sum_i p(a_i)H(p_R|a_i) \tag{16}$$

Das sind die Hörerkosten, also die mittlere Unsicherheit, die verbleibt, wenn der Hörer einen Ausdruck gehört hat.

$$KH = H(p_R|A) = \sum_i p(a_i)H(p_R|a_i) \tag{17}$$

Beispiel 1: sei $p(r_j|a_i) \in \{0, 1\}$ f.a. i, j (es gibt keine Ambiguität). Dann ist entweder $p(r_j|a_i) = 0$ oder $log\ p(r_j|a_i) = 0$.
Also: $H(p_R|A) = 0$, keine Hörerkosten! Unambiguität liefert uns also die untere Schranke!

Beispiel 2: es ist bekannt dass $H(p_R|A) \leq H(p_R)$ (es gibt keine negative Information). Also nimm an, jedes Zeichen kann alles bedeuten. Dann gilt:

$$p(r_j|a_i) = p(r_j) \tag{18}$$

(konditional unabhängig), also

$$H(p_R|A) = H(p_R) \tag{19}$$

Wir haben also wieder eine obere und untere Schranke für die Kosten.

Die Fragestellung

Die Frage lautet: gegeben zwei Mengen A, R, welches ist die Bedeutungsmatrix **B**, so dass der Term

$$\lambda HS + (1 - \lambda)HK \tag{20}$$

minimal wird?

Das ist schwierige Frage, wir müssen berücksichtigen:

- es gibt Synonymie: mehrere Zeichen bezeichnen dieselben Referenten
- es gibt Ambiguität: ein Zeichen bezeichnet mehrere Referenten

und beide interagieren jeweils mit der Gesamtheit - jede einzelne Verschiebung in der Matrix kann **sämtliche** Wahrscheinlichkeiten ändern!

Das Problem

Auch numerisch ist das nicht banal:

- Die Anzahl der Bedeutungsmatrizen **B** für A, R ist endlich, aber:
- Es gibt $2^{|A|\cdot|R|}$ solcher Matrizen, von denen die allermeisten unseren Ansprüchen genügen!
- Also: bereits bei $|A|, |R| > 7$ ist eine *brute force* Suche bereits aussichtslos.

Der heuristische Ansatz

FS nähern sich dem Problem heuristisch:

1. Startkonfiguration ist die Diagonalmatrix
2. Es wird (zufällig) ein Eintrag geändert, so dass die Matrix wohlgeformt bleibt
3. Es wird geprüft, ob

$$\lambda HS' + (1-\lambda)HK' < \lambda HS + (1-\lambda)HK \tag{21}$$

4. Falls nein, wird die alte Matrix erneut modifiziert, falls ja, wird die neue Matrix weiter modifiziert
5. Es werden Schritte 2-4 wiederholt, bis gilt: wir haben die Matrix $2 \cdot |A| \cdot |R|$ mal modifiziert, und konnten die Kosten nicht reduzieren. Dann wird terminiert.

Dieser Ansatz ist also komplett uninformiert, und die Stop-Bedingung ist auch relativ willkürlich!

Evaluation

Die Frage ist: wie evaluiert man jetzt die Ergebnisse? FS nutzen verschiedene Methoden. Die interessanteste ist die **gegenseitige Information**:

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{22}$$

Man misst damit, wieviel Unsicherheit die Kenntnis von Y von X *abzieht*. Der Begriff ist interessanterweise symmetrisch, dass folgt aus den bekannten Symmetrien der Wahrscheinlichkeitstheorie.

Evaluation

Interessant ist: wenn wir den Graphen betrachten für

$$x \mapsto \lambda,\ y \mapsto I(p_A, p_R)$$

basierend auf der optimalen Bedeutungsmatrix (für das entsprechende λ), dann ergibt sich eine **Sigmoid-Kurve**!
Das bedeutet:

- für $\lambda \approx 0.5$ (genauer: 0.4x) ist $I(p_A, p_R) \approx 0.5$

Sprich: ein Zeichen halbiert die Entropie der Referenten (und umgekehrt).

Evaluation

Interessanterweise gibt es dieselbe Verteilung für die **relative Größe des effektiven Lexikons**:

$$L = \frac{|i| \sum_j B_{ij} > 0}{n} \tag{23}$$

Also die Proportion der Ausdrücke, die einen Referenten bezeichnen!
Sprich: für $\lambda \approx 0.5$ ist optimalerweise jeder zweite Ausdruck referierend, es gibt doppelt so viele Referenten wie Ausdrücke!

Evaluation

Noch eine interessante Beobachtung: die optimale Matrix $\hat{B}_\lambda$ legt immer eine Wahrscheinlichkeitsverteilung p_A fest. Hier gilt:

Zipf

Für $\lambda \approx 0.4x$ generiert $\hat{B}_\lambda$ ein p_A mit Zipfverteilung, d.h.

$$p_A(a) \propto rang(a)^{-1} \tag{24}$$

So wie man das empirisch in *allen* natürlichen Sprachen findet!
(Für größere/kleinere λ gilt das übrigens nicht).

Also: eine ganze Reihe sehr interessanter Ergebnisse finden wir hier!

Es gibt aber auch Probleme.

Wiederholung: die Annahme

Das FSZ-Szenario

Nimm an, S will einen Sachverhalt P kommunizieren, und H will ihn wissen.

- S will P möglichst kompakt machen (mit wenig Worten sagen)
- H will, dass P mit möglichst unambig kommuniziert wird (sonst muss er überlegen)

Dieses Szenario ist etwas problematisch:

Probleme beim FSZ-Szenario

Es ist normalerweise nicht der Fall, dass es diese perfekte Antisymmetrie von Sprecher und Hörer gibt:

- Falls der Sprecher etwas will, will er das möglichst deutlich artikulieren
- Falls der Hörer kein Interesse hat, will er möglichst wenig hören.

Das lässt sich leicht an einigen Beispielen darstellen:

Probleme beim FSZ-Szenario

Bsp. 1

Ich möchte dass meine Kinder aufräumen. Ich gebe dafür möglichst präzise Anweisungen (das ist wichtig, hab ich mal gelesen). Meine Kinder versuchen mich möglichst zu ignorieren.

Bsp. 2

Sie treffen einen alten Bekannten, der Ihnen langwierig vom letzten Urlaub erzählt. Sie haben noch was vor und denken sich im stillen: jetzt fass Dich halt mal kurz!

Probleme beim FSZ-Szenario

Bsp. 3

Ich bekomme im Unterricht eine Frage gestellt, auf die ich die Antwort nicht wirklich weiß, aber wissen sollte. Ich antworte möglichst vage. Sie wollen es auch nicht so genau wissen, und sind damit zufrieden.

Bsp. 4

Die Sprache von Kleinkindern hat zu einem gewissen Punkt eine Ähnlichkeit mit L_A. Man kann sehr gut beobachten, wie die Differenzierung verläuft über Konzepten, die das Kind deutlich ausdrücken will (d.h.: Dinge die es will).
Anders gesagt: das Kind verlässt L_A, weil es Dinge deutlich sagen will, und nicht weil es Dinge deutlich hören will.

Probleme beim FSZ-Szenario

Was man hier sieht:

- Es ist zu simpel anzunehmen, dass Sprecher sich per se kurz fassen wollen, und Hörer gern lange hören.
- Das hängt vielmehr von der speziellen Interessenlage ab!

Heißt das, die ganze Arbeit war wertlos? Ich persönlich glaube nicht, man muss die beiden Kostenfunktionen nur anders interpretieren.

Statt Sprecherinteressen und Hörerinteresse sollte man *abstraktere Konzepte* setzen!

Der nächste Artikel

Piantadosi, Tily, Gibson: The communicative function of ambiguity in language (Cognitition 2012)

Die Hypothese

PTG nehmen die FSZ-Hypothese auf, und sehen ihre große Schwachstelle.

Sie versuchen, anderweitig Gründe für die Ambiguität natürlicher Sprache zu finden.

Dabei haben sie zwei (verwandte) Hypothesen.

Hypothese 1

Jedes Kommunikationssystem, das auf Kontextwissen basiert, ist effizienter, wenn es (für sich betrachtet) ambig ist.

Hypothese 2

Jedes effiziente Kommunikationssystem, das auf Kontextwissen basiert, wird einfache Codewörter stärker mehrfach belegen als komplexe Codewörter. (Wobei komplex ein komplexer Begriff ist, Länge, Phonotaktik, etc.)

Zu Hypothese 1

Es gilt immer und allgemein:

$$H(X) \geq H(X|Y) \tag{25}$$

oder noch allgemeiner:

$$H(X|Y_1) \geq H(X|Y_1, Y_2) \tag{26}$$

Anders gesagt: jede Information die wir bekommen, reduziert unsere Entropie (sprich Unsicherheit).

Anders gesagt: es gibt keine negative Information.

Zu Hypothese 1

Sei nun

$$H(p_R|A) = \sum_i p(a_i)H(p_R|a_i) \tag{27}$$

die Restentropie der Referenten, gegeben jeweils einen linguistischen Ausdruck.

$$H(p_R|A, C) = \sum_{i,c \in C} p(a_i, c)H(p_R|a_i, c) \tag{28}$$

ist dann die Restentropie der Bedeutung gegeben Ausdruck *und* Kontext. Es gilt nun natürlich: Kontext ist relevant für Ambiguität/Disambiguierung, also:

$$H(p_R|A) > H(p_R|A, C) \tag{29}$$

Zu Hypothese 1

PTG Argument 1

- Ein effizientes Kommunikationssystem wird versuchen, $H(p_R|A, C)$ niedrig zu halten, während die Anzahl/Länge der Ausdrücke $a \in A$ minimiert wird.
- Das kann natürlich gemacht werden, indem kurze Ausdrücke mehrfach belegt werden.
- Daher *erwarten* wir, dass

$$H(p_R|A) \cong \text{Ambiguität im System}$$

  für sich betrachtet vorhanden ist.
- Denn falls $H(p_R|A) \approx 0$, dann wäre $H(p_R|A, C) \approx 0 \approx H(p_R|A)$, was bedeutet: **Kontextinformation wird nicht genutzt**!

Hypothese 1

Soweit so gut, da kann man sicher nichts sagen. Das Argument ist sicher korrekt, aber deswegen auch ein bisschen lahm (offensichtlich).

V.a. ist es nicht empirisch, sondern eine offensichtliche Beobachtung über Wahrscheinlichkeitsverteilungen!

Kommen wir also zu Hypothese 2!

Hypothese 2

Jedes effiziente Kommunikationssystem, das auf Kontextwissen basiert, wird einfache Codewörter stärker mehrfach belegen als komplexe Codewörter. (Wobei komplex ein komplexer Begriff ist, Länge, Phonotaktik, etc.)

Das ist empirisch überprüfbar, und das ist was PTG machen!

Die Große Frage: wie definiert man die Komplexität eines Wortes?

PTG nutzen den Begriff *effort*; den brechen Sie runter auf drei messbare Faktoren:

1. Länge (je länger desto schwieriger)
2. Häufigkeit (je häufiger desto weniger schwierig)
3. Phonotaktische Wahrscheinlichkeit (Markov-basiert)

Die Liste ist für die Autoren nicht vollständig, und der Begriff des *effort* bleibt etwas vage; es geht nur darum, Faktoren zu haben, die mit effort korrelieren und messbar sind.

Evaluation

Die drei Faktoren werden nun auf zwei Arten von Ambiguität getesten:

1. Homophonie (klar abgetrennte Bedeutungen pro phonologische Form)
2. Polysemie (verwandte Bedeutung pro Form)

(Die Abtrennung ist relativ willkürlich und wird auch nicht weiter vertieft)

Evaluation

Wir bekommen also 6 Tafeln, die überall, die alle sehr grob dasselbe Muster zeigen:

Ein leichtes Fallen der Ambiguität mit steigenden *effort*-Parameter! (Siehe Grafiken)

Wir haben also:

$$\text{Ambiguität} \propto \text{effort}^{-1} \tag{30}$$

Diskussion

Die Ergebnisse sind mit Vorsicht zu geniessen, v.a. wegen *verdeckten Korrelationen*. Z.B.:

Verdeckte Korrelationen

- Länge korreliert mit Häufigkeit (empirisch)
- Phonotaktische Wahrscheinlichkeit korreliert mit Häufigkeit (per Definition: die Wahrscheinlichkeit ist geschätzt aus der Häufigkeit)

Demensprechend musss man vorsichtig sein, wie die Korrelationen kausal zu interpretieren!

Diskussion

Dennoch erweisen sich die Phänomene als sehr robust (und wenig überraschend)

Von daher gibt es hier also solide Evidenz dafür, dass Ambiguität ein Mittel ist, das Kommunikation in menschlicher Sprache *effizient* gestaltet!