SEMPRIA-Whitepaper:

Bedeutungssuche – tiefe semantische Suchmaschinen

Dr. Sven Hartrumpf, Kim Stedwell
SEMPRIA GmbH
Grafenberger Allee 277–287
40237 Düsseldorf
https://www.sempria.de/

2018-09-25

Inhaltsverzeichnis

1 Einleitung

Sprache und Bedeutung scheinen auf den ersten Blick untrennbar miteinander verknüpft. Selten macht es Sinn, beides unabhängig voneinander zu betrachten. Mit der digitalen Sprachverarbeitung gehen jedoch einige Phänomene einher, die zeigen, dass sich Sprache und Bedeutung auf verschiedenen Ebenen und als voneinander getrennt betrachten lassen. Ein einfaches Beispiel dafür ist die Dokumentensuche. Blitzschnell und fehlerlos kann diese jedes Vorkommen des Wortes *Kohl* in einem Dokument wiederfinden und das, ohne auch nur die geringste Ahnung davon zu haben, was das Wort *Kohl* überhaupt bedeutet.

Nicht wenige namhafte Websuchmaschinen funktionieren nach einem sehr ähnlichen Prinzip. Zugegeben, diese Suchmaschinen sind mit dem Lauf der Zeit durchaus raffinierter und besser geworden, jedoch bleibt ihnen die Ebene der Bedeutung weiterhin verschlossen. Noch immer verstehen diese Suchmaschinen nicht, was sie eigentlich suchen. Es hat beinahe den Anschein, als sei dies für eine gute Suche nicht notwendig. Das ist verwunderlich, wenn man sich die scheinbar enge Verbindung von Sprache und Bedeutung wieder ins Gedächtnis ruft. Und es dauert nicht lange, bis erste Probleme auftauchen, bei denen die Betrachtung der Zeichenkette *Kohl* allein nicht mehr ausreicht und die Bedeutungsebene und auch der Kontext miteinbezogen werden müssen, um eine genauere Suche zu ermöglichen. So kann beispielsweise die Zeichenkette *Kohl* zum einen für das Gemüse stehen (Grünkohl, Weißkohl, Chinakohl etc.) zum anderen kann jedoch auch der kürzlich verstorbene Altkanzler Helmut Kohl gemeint sein. Das ist problematisch, da der Suchende in der Regel nur an einer der zwei möglichen Bedeutungen interessiert sein dürfte.

Im Folgenden soll nun gezeigt werden, wie sich die Bedeutungsebene für Suchmaschinen nutzbar machen lässt, was das für die Suchmaschine (und ihre Nutzer) bedeutet und inwiefern dies zu einer besseren, vollständigeren und genaueren Suche führt.

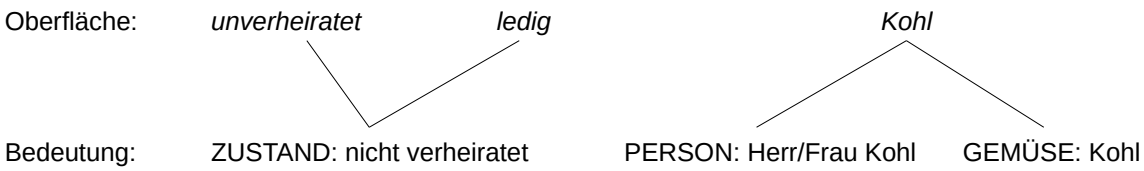

Abbildung 1: Zentrale Unterscheidung von Oberfläche (Wortformen) und Bedeutung (Konzepte).

2 Von der Oberfläche zur Bedeutung

Um das Verhältnis der sprachlichen Oberfläche und der Bedeutung noch ein wenig zu verdeutlichen, sollen zunächst ein paar kennzeichnende Fälle genannt werden. Sprachliche Oberfläche (im Folgenden einfach Oberfläche) bezeichnet in diesem Fall die vorliegenden Zeichenketten. Um auf das Beispiel aus Abschnitt 1 zurückzukommen: Die Zeichenkette *Kohl* ist die Oberfläche. Die Bedeutung kann nun entweder der besagte Altkanzler sein oder aber das Gemüse. In diesem Fall gibt es also mehrere Bedeutungen zur gleichen Oberfläche. Dazu im Gegensatz stehen Synonyme. So unterscheiden sich die Worte *ledig* und *unverheiratet* in der Oberfläche, haben dabei aber dieselbe Bedeutung (s. Abbildung 1).

Die semantische Suchmaschine nutzt eine interne Bedeutungsrepräsentation, um Bedeutungen abbilden und vergleichen zu können. Die genauere Funktionsweise der Bedeutungsrepräsentation würde den Rahmen des vorliegenden Artikels sprengen und soll deshalb ausgeklammert werden. Bei Interesse sei an dieser Stelle jedoch auf ein anderes Whitepaper (Hartrumpf and Helbig, 2016) verwiesen, in dem näher darauf eingegangen wird.

Schon bei einzelnen Wörtern zeigt sich, dass sich die beiden Ebenen, Oberfläche und Bedeutung, unterschiedlich verhalten können. Dies spitzt sich weiter zu, wenn man Phrasen oder ganze Sätze betrachtet. Als Beispiel sollen folgende Sätze dienen:

1. *X übt Kritik an Y.*
2. *X kritisiert Y.*
3. *Y kritisiert X.*

Betrachtet man allein die Oberfläche dieser Sätze und lässt die Bedeutung völlig außen vor (was für den Menschen nahezu unmöglich ist), so könnte fälschlicherweise angenommen werden, dass sich 2. und 3. ähnlicher sind als die Sätze 1. und 2. Erst wenn auch die Bedeutung der Sätze untersucht wird, lässt sich feststellen, dass sich die Sätze 1. und 2. ähnlich sind (denn sie gleichen sich in der Bedeutung) und die Sätze 2. und 3. etwas grundsätzlich anderes aussagen. Diese Bedeutungsunterschied kann nur erfasst werden, wenn die Worte *Kritik* und *kritisieren* sowie deren Zusammenhang richtig verstanden wird. Zum Wort *kritisieren* (und auch zum Wort *Kritik*) gehören vereinfacht gesagt zwei *Mitspieler*. Damit ist zum einen der/die Kritisierende und zum anderen der/die Kritisierte gemeint. Diese Rollen sind nicht austauschbar - *Roger Ebert kritisiert Michael Bay* ist nicht dasselbe wie *Michael Bay*

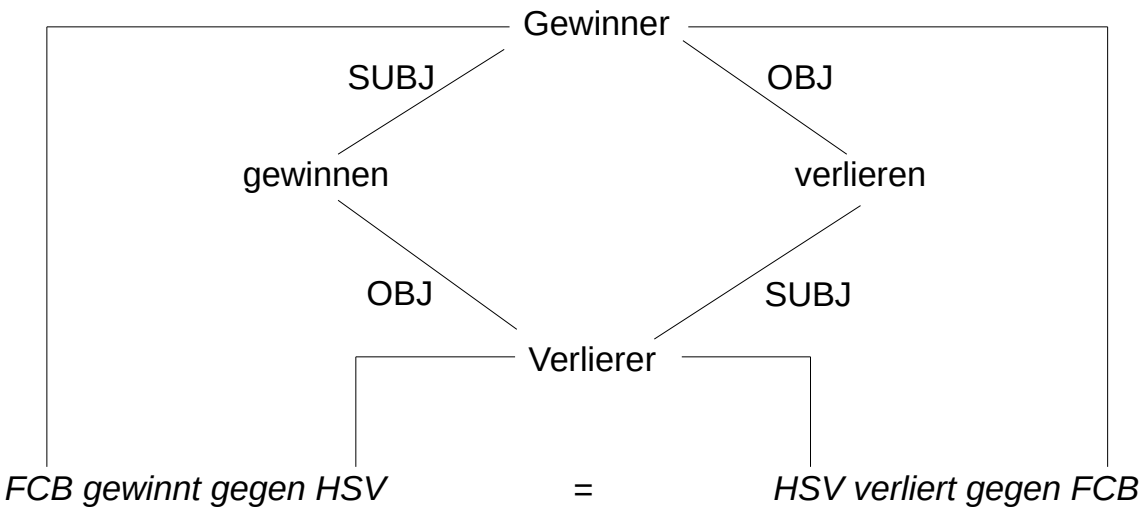

Abbildung 2: Zusammenhang von *gewinnen*/*Gewinner* und *verlieren*/*Verlierer* (zur Illustration vereinfacht).

kritisiert Roger Ebert, denn die Rollen sind vertauscht. Genauso wichtig ist zu verstehen, dass die Prädikate *kritisieren* und *Kritik üben* die gleiche Bedeutung (und die gleichen Mitspieler) haben. Dementsprechend bedeutet die Sätze *Roger Ebert kritisiert Michael Bay* und *Roger Ebert übt Kritik an Michael Bay* dasselbe, denn die Prädikate bedeuten dasselbe und auch die Rollen der Mitspieler sind richtig besetzt. Ein Beispiel von komplexen Beziehungen zwischen verschiedenen Verb-Konzepten (*gewinnen* und *verlieren*) und verwandten Wörtern illustriert Abbildung 2.

Das Verhältnis von Oberfläche und Bedeutung sollte keinesfalls als Problem verstanden werden, wie die genannten Beispiel womöglich suggerieren könnten. Ein Problem ist nur, dass die herkömmlichen Suchmaschinen die Bedeutungsebene ignorieren und somit ein großes Potential zur besseren Suche verschenken. Eine tiefe semantische Suche nutzt eben dieses Potential. Was das für die Suche bedeutet, wird in den nächsten Abschnitten behandelt.

3 Vollständigkeit der Suchergebnisse

Ein Kriterium, an dem die Qualität einer Suchmaschine gemessen werden kann, ist die Vollständigkeit der Suchergebnisse. Wird ein Archiv mit einem bestimmten Begriff durchsucht, so ist die Suche vollständig, wenn sämtliche Dokumente, die für den gesuchten Begriff relevant sind, gefunden werden. Nicht immer können alle relevanten Dokumente gefunden werden, deshalb gilt vereinfacht gesagt: Je mehr relevante Dokumente gefunden werden, desto besser ist die Suche.

Anhand eines Beispiels soll nun gezeigt werden, inwiefern die verschiedenen Suchmaschinen (traditionelle und tiefe semantische) zu unterschiedlichen Ergebnissen kommen können. Zu diesem Zweck wird von einem Zeitungsarchiv ausgegangen. Dieses Archiv wird mit der Suchanfrage *Donald Trump unterschreibt Einwanderungsgesetz* durchsucht. Es existieren vier relevante Treffer im Archiv:

1. *US-Präsident Donald Trump unterschreibt Einwanderungsgesetz. Die gilt als umstritten.*

2. *Donald Trump unterschreibt das als umstritten geltende neue Einwanderungsgesetz.*

3. *Soeben wurde ein neues Einwanderungsgesetz von US-Präsident Trump unterschrieben.*

4. *US-Präsident unterzeichnet neues Gesetz zur Einwanderung.*

Zunächst wird die traditionelle Suchmaschine betrachtet, die lediglich die Oberfläche der Suchanfrage untersucht. Satz 1 sollte mit Sicherheit gefunden werden, denn er enthält die Suchanfrage exakt wie sie der Suchmaschine übergeben wurde (Zeichen für Zeichen). Schon bei Satz 2 wird deutlich, dass nicht immer Zeichen für Zeichen gesucht werden kann. Allerdings sollte die herkömmliche Suchmaschine auch bei Satz 2 fündig werden, enthält er doch sämtliche Begriffe, nach denen gesucht wurde. Satz 3 bringt nun weitere Probleme mit sich. Zum einen ist hier noch mehr von *Donald Trump* sondern von *US-Präsident Trump* die Rede. Außerdem taucht das Prädikat hier in einer Passiv-Konstruktion auf und damit als *unterschrieben* und nicht als *unterschreibt*. Rein oberflächlich bestehen also nun schon mehrere Unterschiede zwischen der Suchanfrage und dem Dokument. Ob Satz 3 von der herkömmlichen Suchmaschine gefunden wird, ist fraglich. Im letzten Satz sind die Unterschiede so groß, dass die herkömmliche Suchmaschine dieses Dokument mit großer Sicherheit übersieht. Keines der Wörter aus der Suchanfrage kommt im Dokument vor. Statt *Donald Trump* steht dort nur *US-Präsident*, statt *unterschreibt* steht dort *unterzeichnet* und auch *Einwanderungsgesetz* wurde durch *neues Gesetz zur Einwanderung* realisiert.

Die tiefe semantische Suche schlägt sich da deutlich besser. Der erste Satz, der mit der Suchanfrage identisch ist, stellt natürlich auch für die semantische Suche kein Problem dar. Der zweite Satz enthält, wie auch die Suchanfrage, das Prädikat *unterschreiben*. Auch die Rollen der Mitspieler (der/die Unterschreibende und das Unterschriebene) sind gleich besetzt. Der dritte Satz wird auf ganz ähnliche Weise analysiert: Das Prädikat ist noch immer das gleiche und, auch wenn die grammatikalische Realisierung hier etwas anders ist (eine Passiv-Konstruktion), sind die Rollen des Prädikates weiterhin gleich besetzt. Dies klappt auch deshalb, weil die Begriffe *Donald Trump* und *US-Präsident Trump* für die tiefe semantische Suchmaschine Synonyme sind. Ähnlich ist das auch bei Satz 4. *US-Präsident* wird ebenfalls als synonym für *Donald Trump* verstanden (zumindest solange dieser das Amt ausübt). Auch *unterzeichnen* wird als Synonym von *unterschreiben* erkannt. Interessant wird es auch bei den Begriffen *Einwanderungsgesetz* und *Gesetz zur Einwanderung*. Der Vergleich dieser Begriffe funktioniert ebenfalls anhand seiner *Mitspieler*. So füllt das Wort *Einwanderung* eine bestimmte Rolle, die das Wort Gesetz vorsieht. Egal ob dies nun in Form von *Einwanderungsgesetz* oder *Gesetz zur Einwanderung* realisiert wird, die tiefe semantische Suche erkennt, dass die *Einwanderung* in beiden Fällen dieselbe Rolle füllt. Damit sind alle Teile des Satzes erfolgreich mit der Suchanfrage abgeglichen und auch die Bedeutung des ganzen Satzes stimmt mit der Bedeutung der Suchanfrage überein.

Aus diesem Beispiel geht hervor, dass die herkömmliche Suchmaschine darauf angewiesen ist, dass die Suchanfrage dem Text des relevanten Dokuments gleicht oder zumindest sehr ähnlich ist. Bedenkt man aber nun die unzähligen sprachlichen Möglichkeiten, mit denen ein und derselbe Sachverhalt ausgedrückt werden kann, so sollte offensichtlich sein, dass dieser Ansatz sehr begrenzt ist. Die tiefe semantische Suchmaschine hingegen vergleicht nicht die Oberflächen, sondern Bedeutungen. Wie diese Bedeutungen letztendlich oberflächlich aussehen, spielt dabei nur eine untergeordnete Rolle.

Allerdings lässt sich die Qualität der Suchmaschine nicht allein an der Vollständigkeit messen, wie Abschnitt 4 zeigen wird.

4 Genauigkeit der Suchergebnisse

Besonders bei größeren Archiven ist es auch für die herkömmliche Suchmaschine kein Problem, eine große Anzahl Treffer zu liefern. Das Problem besteht eher darin, dass häufig ein großer Teil der angezeigten Treffer irrelevant ist, diese werden im Folgenden als *irrelevante Treffer* bezeichnet. Ein irrelevanter Treffer ist ein Dokument, dass von der Suchmaschine fälschlicherweise für relevant gehalten wird, obwohl es mit der Suchanfrage nichts zu tun hat oder gar mit diesem im Widerspruch steht. Das zeigt, warum Vollständigkeit allein nicht ausreicht, um die Qualität einer Suchmaschine zu messen. Denn was nützt es schon, wenn eine Suchmaschine sämtliche relevanten Treffer findet, wenn sich diese in einer großen Anzahl irrelevanter Treffer verlieren? Die Suche muss nicht nur vollständig, sondern auch genau sein. In diesem Fall bedeutet Genauigkeit die Abwesenheit von irrelevanten Treffern. Oder anders: Je weniger irrelevante Treffer eine Suchmaschine liefert, desto genauer ist sie.

Ähnlich wie im Abschnitt 3 soll anhand eines Beispiels verdeutlicht werden, wie die tiefe semantische Suche im Vergleich zur herkömmlichen Suche mit irrelevanten Treffern umgeht, oder vielmehr, wie diese vermieden werden sollen. Wieder soll von einem Archiv ausgegangen werden. Gesucht wird mit *Sieg des FC Bayern*. Folgende Dokumente, die von der traditionellen Suchmaschine für relevant gehalten werden könnten, sind im Archiv enthalten.

1. *Die Fans feiern den Sieg des HSV gegen den FC Bayern.*
2. *So wurde der Sieg des FC Bayern in letzter Sekunde verhindert.*
3. *Der Architekt Sieg entwirft nun doch nicht das neue Stadion des FC Bayern.*

Warum die traditionelle Suchmaschine diese Artikel für relevant hält, sollte recht schnell ersichtlich sein. Alle drei Sätze enthalten sämtliche Wörter, die auch in der Suchanfrage enthalten sind. Dass diese Sätze irrelevant sind, geht aus ihrer Oberfläche nicht hervor. Dazu muss die Bedeutung herangezogen werden.

Die tiefe semantische Suche ist hingegen in der Lage, die Bedeutung der Sätze mit der der Suchanfrage zu vergleichen. Beim ersten Satz kommen dabei wieder die

bereits mehrfach erwähnten Mitspieler zum Tragen. Das Wort *Sieg* hat in diesem Sinne zwei Mitspieler (den *Sieger* und den *Besiegten*). So erkennt die Suchmaschine, dass in Satz 1 nicht von einem Sieg des FC Bayern, sondern von einem Sieg des HSV die Rede ist. Damit ist dieses Dokument irrelevant und die semantische Suche kann es ausschließen. Auch beim zweiten Satz erkennt die tiefe semantische Suche, dass dort eben nicht von einem Sieg die Rede ist, schließlich wurde dieser *in letzter Sekunde verhindert*. In Satz 3 erkennt die tiefe semantische Suche, dass es sich beim Vorkommen des Wortes *Sieg* in diesem Fall um einen Namen handelt. So kann auch der dritte Satz erfolgreich ausgeschlossen werden.

Die Genauigkeit der Suche darf nicht unterschätzt werden. Muss der Nutzer selbst in einer Vielzahl von irrelevanten Treffern nach den relevanten Treffern suchen, so tut er genau das, was eigentlich Aufgabe der Suchmaschine sein sollte.

5 Fazit

In Abschnitt 3 wurde die Wichtigkeit einer vollständigen Suche beleuchtet. Eine gute Suchmaschine findet möglichst viele oder sogar alle relevanten Dokumente zu einer Suchanfrage. Abschnitt 4 hat gezeigt, dass Vollständigkeit allein nicht ausreicht um die Qualität einer Suchmaschine zu ermessen. Selbst wenn eine Suchmaschine alle relevanten Dokumente finden kann, so ist nichts damit gewonnen, wenn diese von irrelevanten Treffern in doppelter Menge überschattet werden. Für eine gute Suche muss also Folgendes angestrebt werden: hohe Vollständigkeit bei bestmöglicher Genauigkeit.

Abschnitt 3 und Abschnitt 4 haben gezeigt, dass die semantische Suchmaschine der traditionellen Suchmaschine überlegen ist, bei der Vollständigkeit und auch bei der Genauigkeit. Damit liefert der semantische Ansatz eine hochwertigere, bessere und komfortablere Suche.

Es sollte erwähnt werden, dass der Aufwand einer semantischen Suche größer ist als der einer traditionellen Suchmaschine. Um auf Bedeutungsebene suchen zu können, muss die Suchmaschine Bedeutungen verstehen. Dazu braucht es unter anderem ein stetig aktualisiertes Lexikon, Grammatikverständnis und ein umfangreiches Faktenwissen. Ebenfalls wichtig ist, dass die traditionelle Suchmaschine sich über die Zeit weiterentwickelt hat, so kommt beispielsweise gelegentlich Synonymwissen zum Einsatz. Dies ist jedoch nicht mehr als ein kleiner Zusatz zu einem Ansatz zu verstehen, der mittlerweile seine Grenzen erreicht. Damit die traditionelle Suchmaschine auch weiterhin den Ansprüchen genügen kann, ist eine Vielzahl von Erweiterungen und Umstellungen erforderlich. Und es sollte darüber nachgedacht werden, ob der damit verbundene Mehraufwand nicht vielleicht sogar größer ist, als der des bedeutungsorientierten Ansatzes der semantischen Suche. Außerdem besteht der Mehraufwand der semantischen Suche besonders darin, zunächst ein Lexikon und eine Grammatik aufzubauen. Haben diese erst einmal einen gewissen Umfang erreicht, so ist der Ansatz der tiefen semantischen Suche für nahezu jede Umgebung gerüstet.

Literatur

Hartrumpf, Sven and Hermann Helbig (2016). Einführung in die bedeutungsorientierte Suchmaschine SEMPRIA Search. Whitepaper, SEMPRIA GmbH. URL https://www.sempria.de/.