Unsupervised Word-Sense Disambiguation

Christian Wurm

Düsseldorf, 11.5.2023

Einleitung

Zur Erinnerung:

- Supervised learning bedeutet, es gibt annotierte Daten, die die richtige Klassifikation vorgeben
- Unsupervised learning bedeutet, es gibt nur rohe Daten; die Daten, auf denen trainiert wird, sehen genauso aus wie die Daten auf denen später operiert wird.

Das bedeutet:

1. Supervised WSD benötigt Korpora, in denen ambige Worte nach Bedeutung annotiert sind
2. Unsupervised WSD benötigt einfach Korpora

Supervised WSD ist im Prinzip simpel heutzutage, das Problem sind die Daten, die es nur für sehr wenige Sprachen in ausreichender Menge gibt.

Einleitung

1. Supervised WSD benötigt Korpora, in denen ambige Worte nach Bedeutung annotiert sind
2. Unsupervised WSD benötigt einfach Korpora

Supervised WSD ist im Prinzip simpel heutzutage, das Problem sind die Daten, die es nur für sehr wenige Sprachen in ausreichender Menge gibt.

Also:

Unsupervised WSD ist nach wie vor eine sehr wichtige Aufgabe, denn alle Datenmengen, die wir mittlerweile haben, sind unannotiert, und die neuesten Algorithmen, die wir haben, sind supervised ML.

Bootstrapping

Bootstrapping ist ein (sehr allgemeines) Verfahren, wie man von einigen wenigen Beispielen auf viele und letztenendes auf alle Fälle schließt. Die Methode dabei ist folgende:

1. Nimm eine begrenzte Anzahl von Beispielen für Klasse 1 und 2, anhand von bestimmten Merkmalen
2. Nutze diese Beispiele, um neue Merkmale zu generieren
3. Nutze diese Merkmale, um neue Beispiele zu finden
4. Gehe zurück zu 2, es sei denn, wir haben keine neuen Beispiele mehr – dann terminiere.

Bootstrapping basiert also auf der Unterscheidung von Objekte und Merkmalen, bekannt als **Extension** und **Intension**.

Bootstrapping bei Yarovski

Yarovski macht sich diese Methode zunutze:

- Die "Objekte" sind die Lesarten eines gegebenen Wortes
- Die "Merkmale" sind die Worte, mit denen es zusammen vorkommt, entweder im selben Wortfenster, Satz oder Dokument.

Den Anfang machen sehr wenige (nur eines) von Hand ausgewählte Kontextworte.

Die Methode

Wir nehmen also ein Wort w und unterscheiden Lesarten $b_1(w), ..., b_n(w)$. Ebenfalls nehmen wir für jede Lesart ein Kontextwort:

- $b_1(w)$ korrelliert mit v_1
- ...
- $b_n(w)$ korrelliert mit v_n

Wir müssen nun Unterscheiden zwischen Worten und Kontexten.

Die Methode

Wir definieren die Intension $I_0(b_i(w)) = \{v_i\}$

$I_{n+1}(b_i(w)) = \{v :$

1. v steht im Kontext von einer Okkurrenz von w
2. diese Okkurrenz von w steht im Kontext von $v' \in I_n(b_i(w))\}$
3. falls u, v im Kontext einer Okkurrenz von w stehen, $i \neq j$, dann $u \notin I_n(b_j(w))$

$\}$

Wir verlangen hier also Eindeutigkeit.

Die Methode

$I_{n+1}(b_i(w)) = \{v :$

1. v steht im Kontext von einer Okkurrenz von w
2. diese Okkurrenz von w steht im Kontext von $v' \in I_n(b_i(w))\}$
3. falls u, v im Kontext einer Okkurrenz von w stehen, $i \neq j$, dann $u \notin I_n(b_j(w))$

$\}$

Hier kann man verschiedene Schrauben drehen: z.B. bestimmte Kookkurrenz-Wahrscheinlichkeiten anstelle von Eindeutigkeit, z.B. die berühmten 95%

3. in 95% aller Fälle, in denen v im Kontext einer Okkurrenz von w steht, gibt es kein u in demselben Kontext mit $i \neq j$ und $u \in I_n(b_j(w))$

Die Methode

Diese Methode wird, gegeben einen endlichen Datensatz irgendwann konvergieren; spricht:

Konvergenz

Es gibt ein $m \in \mathbb{N}$ so dass für alle $1 \leq i \leq n$ gilt:

$$I_m(b_i(w)) = I_{m+1}(b_i(w)) \tag{1}$$

Wir nennen diese Funktion I_∞, denn wir könnten die Methode genauso gut unendlich iterieren.

Wir definieren den Kontext einer Okkurrenz eines Wortes der Einfachheit halber $C(w)$.

Die Methode

Damit bekommen wir einen einfachen Klassifikator f_{WSD}:

$f_{WSD}(w) = b_i(w)$ gdw.

1. $I_\infty(b_i(w)) \cap C(w) \neq \emptyset$
2. falls $j \neq i$, $I_\infty(b_j(w)) \cap C(w) = \emptyset$

Alternativ: $f_{WSD}(w) = b_i(w)$ gdw. f.a. $j \neq i$ gilt:

$$|I_\infty(b_i(w)) \cap C(w)| > |I_\infty(b_j(w)) \cap C(w)| \tag{2}$$

Die Methode

Damit bekommen wir einen einfachen Klassifikator f_{WSD}:

Dadurch, dass die Intensionen disjunkt sind, ist die Zuordnung eindeutig; es kann aber durchaus passieren, dass eine Zuordnung nicht möglich ist.

$f_{WSD}(w) = b_i(w)$ gdw.

1. $I_\infty(b_i(w)) \cap C(w) \neq \emptyset$
2. falls $j \neq i$, $I_\infty(b_j(w)) \cap C(w) = \emptyset$

- Hier sieht man, wie wichtig die Wahl des richtigen Kontextes $C(w)$ ist:
- Je kleiner $C(w)$, desto eher können wir Bedingung 1 nicht erfüllen
- Je größer $C(w)$, desto eher können wir Bedinung 2 nicht erfüllen!

Literatur

- Yarovski: Unsupervised word sense disambiguation rivaling supervised methods