Aufbauseminar Python

Word Sense Disambiguation(wissensbasiert)

Ahmed Albawabiji

Professur für Korpuslinguistik
Studiengang Linguistische Informatik
Friedrich-Alexander-Universität Erlangen-Nürnberg
ahmed.albawabiji@fau.de

Erlangen, 03.02.2017

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Inhaltsverzeichnis

Die Papers:

Ted Pederson(2001). A Decision Tree of Bigrams is an Accurate Predictor of Word Sense, USA.

Satanjeev Banerjee, Ted Pederson(2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, USA.

Begriffserkärung

Word Sense Disambiguation(WSD): Lesartendisambiguierung.
Disambiguieren die Sinne des Wortes.
Die Mehrdeutigkeit eines Wortes in einem bestimmten Zusammenhang eindeutig machen.
Erwünschte Mehrdeutigkeit bei Witzen.
Unerwünschtes Problem beim maschinellen Lernen.
Als Lösung:Eine bestimmte Sense mit Hilfe der Algorithmen festlegen.

Motivation

Warum geht eine Blondine mit einem Wassereimer in die Bank?

Weil sie ihr Konto löschen will.
(1) Konto löschen (to delete)
(2) **mit Wasser** Feuer löschen (to extinguish)

gibt es Mehrdeutigkeit zwischen Menschen?

Person (A): Was kann ich gegen die fliegenden Fliegen tun ?

Lesk

- **Grundidee:** Indikatorwörter für Lesarten werden aus Bedeutungsdefinitonen gewonnen.
- Lexikalisch-semantische Disambiguierung mittels WordNet.
- wird für maschinelles Lernen und andere Zwecke benutzt.
- Wichtigste Indikatoren für Lesart im Kontext sind die Inhaltswörter im Satz = Kontextwörter
- Lesk's Ansatz nutzt Relationen zwischen Synonyms, die WordNet bietet.
- Lesk Algorithmus vergleicht Glossen(Worterklärungen) der Zielwörter. Daher werden die Wörter eingeschränkt.

Lesk

- WordNet: ist ein umfangreiches semantisches Netz von Bedeutungsrelationen für Englisch und verbessert die Genauigkeit userer Disambiguierung.(Semantisch angeordnet, POS steht auch dabei)
- Synsets für Substantive bilden baumförmige Hierarchie = taxonomische Ontologie
- Synonyme sind zusammen gruppiert.
- polysemous. Wenn ein Wort in mehreren synsets auftaucht, wo jede Synset eine mögliche Sense des Wortes repräsentiert.
- Online-Zugang und freie Downloads unter

`http://wordnetweb.princeton.edu/perl/webwn`

vergleichen Glossen zwischen jeden Wortpaaren im Kontextfenster.
Hyperonym(Farbe), Hyponym(Rot). Holonym(Hand), meronym(Finger)
troponym(spazieren von gehen) und Attribute von jedem Wort in Paar.
Einschränkung der Relationen, nur wenn POS bekannt ist. Synset assoziiert mit dem POS.
Der Algorithmus ist unabhängig von Relation und könnte mehrfach durchgeführt werden.

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: bank Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in

Beispiel

- "*The land owner asked the bank to loan him a large sum of money.*"
  - Lesart $bank^{(1)}$: 1 Indikatorwort im Kontext
    "*The **land** owner asked the bank to loan him a large sum of money.*"
  - Lesart $bank^{(2)}$: 2 Indikatorwörter im Kontext
    "*The land owner asked the bank to **loan** him a large sum of **money**.*"
- Lesk-Algorithmus wählt Lesart $bank^{(2)}$ aus
  - Parameter: Lemmatisierung, keine Funktionswörter

Gäbe es bei beiden 2,würde das meist bekannte Wort gewinnen.

Evaluation

Test data von Englisch Lexikal sample SenseEval-2 wurde zur evaluieren benutzt. Es gibt 4,328 Instanzen. Geteilt durch 29 Nomen 29 Verben und 15 Adjektiven.
Accuracy:

- Nomen: 0.322,564 von 1754 richtig
- Verben: 0.249,450 von 1806 richtig
- Adjektiven: 0.469,360 von 768 richtig
- Allgemein: 31.7% wo 1374 von 4328 richtig

Decision Tree of Bigrams

Abstrakt: Korpus basiert auf Ansatz zu WSD, wo decision tree definiert Sense von mehrdeutigem Wort. Evaluation ist durch SenseEval 1998 gemacht worden.jede Sense von einem mehrdeutigen Wort wird zu einem Feature Vektor umgewandelt.
wir haben: POS von Kontextwörtern, syntaktische Eigenschaften vom Zielwort und seinem Satz.
Vorkommen von Bigrams (nacheinander kommende Wörter)

Decision Tree of Bigrams

Ein Baum von Entscheidungen(Sense-Tag) wird für ein bestimmtes Wort gebaut
Jedes Auftreten des Sense- tagged von einem Polynym ist konvertiert zu einem Feature eines Vektors.
Jedes Feature repräsentiert eine Eigenschaft von Umgebungskontext, der als relevant zum Disambiguierungsprocess berücksichtigt ist.
Ein Bigram sind zwei wortketten die im Text vorkommen.
Binäre Features repräsentieren den Kontext von dem Zielwort.
Kontextwörter und sekundäre Kontextwörter werden zu bag of words zusammengefasst (ohne Duplikate) Vergleich mit bag of words der Indikatorwörter einer möglichen Lesart.
wir fügen die meist Information gebenden Features zu einer Baumstruktur hinzu(jedes Feature zu einem Knoten).

	cat	¬cat	totals
big	$n_{11}= 10$	$n_{12}= 20$	$n_{1+}= 30$
¬big	$n_{21}= 40$	$n_{22}= 930$	$n_{2+}= 970$
totals	$n_{+1}=50$	$n_{+2}=950$	$n_{++}=1000$

Figure 1: Representation of Bigram Counts

verschiedene syntaktische Eigenschaften vom Satz und mehrdeutige Wort.
Bigram ist ein lexikales Feature, hilft uns bei WSD.
Es ist aber nicht klar, wie weit!!!
Ein Ziel ist eine klare Abgrenzung der Genauigkeit der Disambiguierung durch die Benutzung Feature sets.

Power Divergence Family(basiert auf Statistik) Likelihood ratio G2 und Pearson's X2 (manchmal mehr zuverlässig) oft unklar, welche besser ist.

$$m_{ij} = \frac{n_{i+} * n_{+j}}{n_{++}}$$

$$G^2 = 2 \sum_{i,j} n_{ij} * \log \frac{n_{ij}}{m_{ij}}$$

$$X^2 = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

$$Dice(w_1, w_2) = \frac{2 * n_{11}}{n_{+1} + n_{1+}}$$

- *"The land owner asked the bank to loan him a lot of money, but was rejected."*
  - Kontextwörter: *land*, *owner*, *loan*, *money*, *reject* (*ask* & *a lot of* zu allgemein → Stoppwörter)
  - Indikatorwörter für Lesart $bank^{(1)}$: *slope*, *land*, *body*, *water*
  - Indikatorwörter für Lesart $bank^{(2)}$: *financial*, *institution*, *accept*, *deposit*, *channel*, *money*, *lend*, *activity*

- Indikatorwörter für *bank*$^{(1)}$:
  - *slope*, *land*, *body*, *water*
- Sekundäre Kontextwörter (optimiert):
  - *land*: "The land on which real estate is located."
    → *land*, *real estate*, *locate*
  - *owner*: "A person who owns something."
    → *owner*, *person*, *own*
  - *loan*: "Give temporarily; let have for a limited time."
    → *loan*, *temporary*, *limited*
  - *money*: "Official currency issued by a national bank."
    → *money*, *official*, *currency*, *issue*, *national*, *bank*
  - *reject*: "Refuse to accept."
    → *reject*, *refuse*, *accept*

- Indikatorwörter für *bank*(1):
  - *body*, *land*, *slope*, *water*
- *Bag of words* der Kontextwörter:
  - *accept*, *bank*, *currency*, *issue*, *land*, *limited*, *loan*, *locate*, *money*, *national*, *official*, *own*, *person*, *real estate*, *refuse*, *reject*, *temporary*
- Dice: $D = 2 \times 1 / (17 + 4) = 2/21 = .0952...$
  - $M = 1$ Übereinstimmung (*land*)
  - $C = 17$ Kontextwörter und sekundäre Kontextwörter
  - $I = 4$ Indikatorwörter

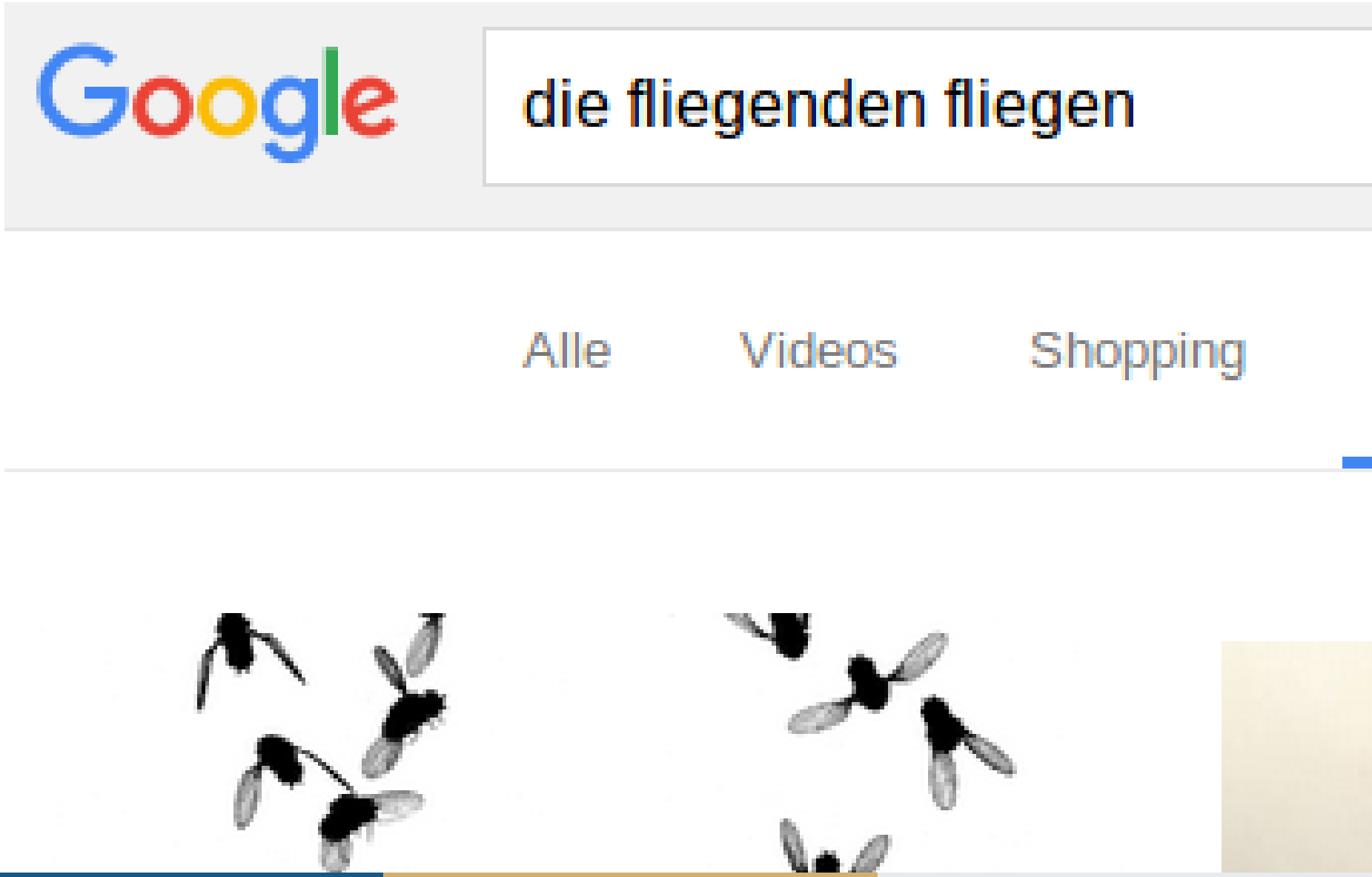
Table 1: Experimental Results

(1) word-pos	(2) test	(3) senses in test	(4) train	(5) maj	(6) best	(7) avg	(8) j48 pow	(9) stump pow	(10) j48 dice	(11) stump dice	(12) naive bayes
accident-n	267	8	227	75.3	87.1	79.6	85.0	77.2	83.9	77.2	83.1
behaviour-n	279	3	994	94.3	92.9	90.2	95.7	95.7	95.7	95.7	93.2
bet-n	274	15	106	18.2	50.7	39.6	41.8	34.5	41.8	34.5	39.3
excess-n	186	8	251	1.1	75.9	63.7	65.1	38.7	60.8	38.7	64.5
float-n	75	12	61	45.3	66.1	45.0	52.0	50.7	52.0	50.7	56.0
giant-n	118	7	355	49.2	67.6	56.6	68.6	59.3	66.1	59.3	70.3
knee-n	251	22	435	48.2	67.4	56.0	71.3	60.2	70.5	60.2	64.1
onion-n	214	4	26	82.7	84.8	75.7	82.7	82.7	82.7	82.7	82.2
promise-n	113	8	845	62.8	75.2	56.9	48.7	63.7	55.8	62.8	78.0
sack-n	82	7	97	50.0	77.1	59.3	80.5	58.5	80.5	58.5	74.4
scrap-n	156	14	27	41.7	51.6	35.1	26.3	16.7	26.3	16.7	26.7
shirt-n	184	8	533	43.5	77.4	59.8	46.7	43.5	51.1	43.5	60.9
amaze-v	70	1	316	0.0	100.0	92.4	58.6	12.9	60.0	12.9	71.4

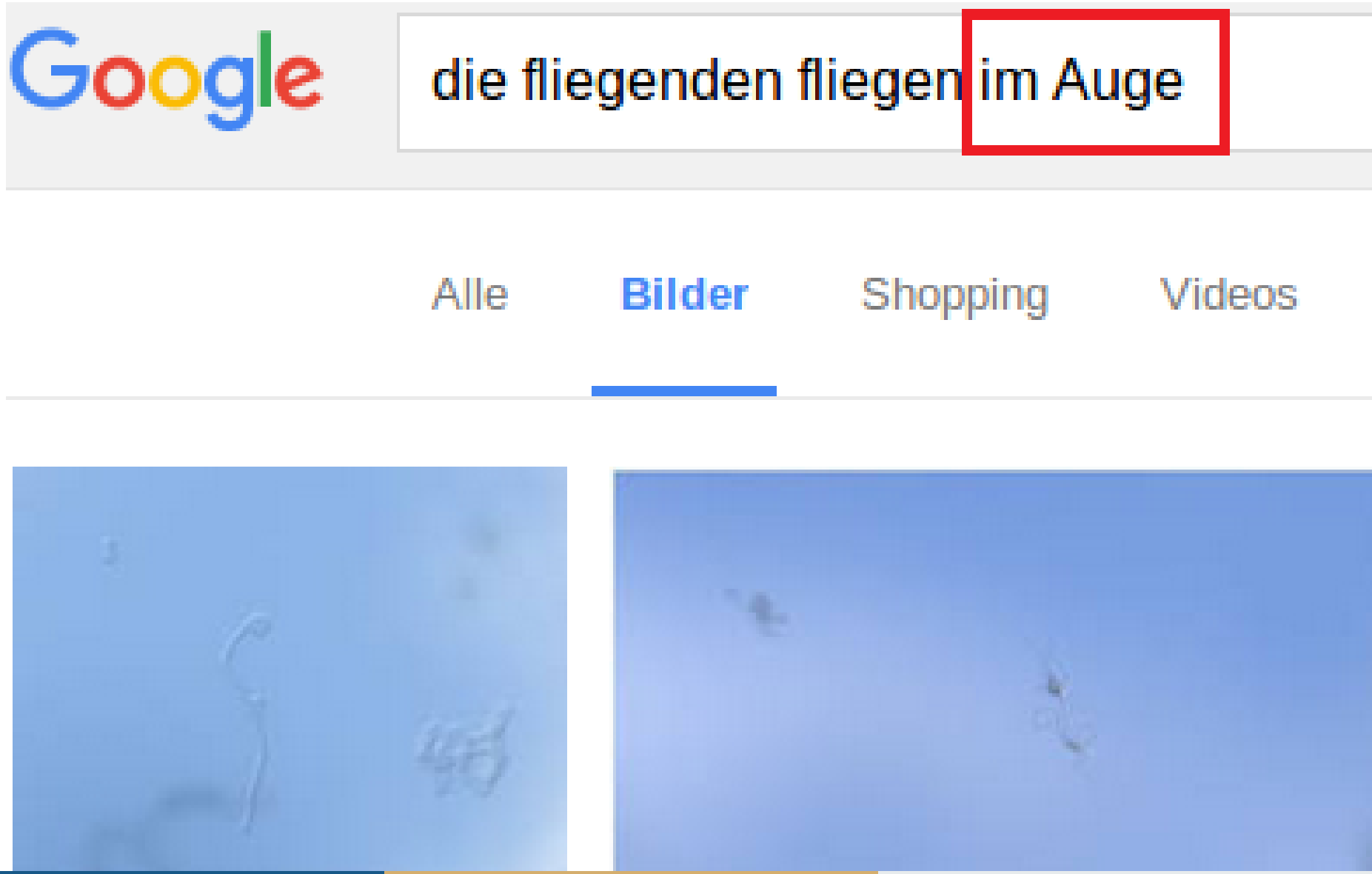
Table 2: Decision Tree and Stump Characteristics

	power divergence			dice coefficient		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
word-pos	stump node	leaf/total	features	stump node	leaf/total	features
accident-n	by accident	8/15	101	by accident	12/23	112
behaviour-n	best behaviour	2/3	100	best behaviour	2/3	104
bet-n	betting shop	20/39	50	betting shop	20/39	50
excess-n	in excess	13/25	104	in excess	11/21	102
float-n	the float	7/13	13	the float	7/13	13
giant-n	the giants	16/31	103	the giants	14/27	78
knee-n	knee injury	23/45	102	knee injury	20/39	104
onion-n	in the	1/1	7	in the	1/1	7
promise-n	promise of	95/189	100	a promising	49/97	107
sack-n	the sack	5/9	31	the sack	5/9	31
scrap-n	scrap of	7/13	8	scrap of	7/13	8
shirt-n	shirt and	38/75	101	shirt and	55/109	101

Was kann ich gegen die fliegenden Fliegen?



Was kann ich gegen die fliegenden Fliegen im Auge tun?



Ted Pederson(2001). A Decision Tree of Bigrams is an Accurate Predictor of Word Sense, USA.
Satanjeev Banerjee, Ted Pederson(2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, USA.
Stefan Evert , Lesartendisambiguierung(WSD)