Automatische Belebtheitsklassifikation im Deutschen

Schriftliche Hausarbeit
für die Bachelorprüfung der Fakultät für Philologie
an der Ruhr-Universität Bochum
(Gemeinsame Prüfungsordnung für das Bachelor/Master-Studium
im Rahmen des 2-Fach-Modells an der RUB vom 7.1.2002)

vorgelegt von

Pagel, Janis Malte

25. August 2015

Erstgutachterin: Prof. Dr. Stefanie Dipper
Zweitgutachter: Prof. Dr. Ralf Klabunde

Inhaltsverzeichnis

Verwendete Abkürzungen

Akk	Akkusativ
Dat	Dativ
fn	False Negative
fp	False Positive
Gen	Genitiv
Mask	Maskulinum
MCC	Matthews Correlation Coefficient
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NP	Noun Phrase / Nominalphrase
P	Precision
PP	Prepositional Phrase / Präpositionalphrase
R	Recall
Sg	Singular
sw	schwach
tp	True Positive
tn	True Negative

1 Einleitung

Belebtheitsklassifikation ist eine im Bereich der maschinellen Sprachverarbeitung oft vernachlässigte Disziplin. Dabei kann die Kenntnis um Belebtheitsmerkmale von Wörtern zu einer Verbesserung der Ergebnisse einer ganzen Reihe von wichtigen Bereichen der NLP wie Anaphernresolution oder automatische Übersetzung führen. Darüber hinaus ist auch Belebtheitsklassifikation für sich genommen ein interessanter Vorgang, da sie Erkenntisse über semantische Zusammenhänge liefern kann und theoretische Ansätze zu Belebtheit in menschlicher Sprache auf die Probe stellt.

In der vorliegenden Arbeit möchte ich eine Methodik vorstellen, die Belebtheit im Deutschen erkennen und korrekt zuordnen soll. Dabei werde ich auf verschiedene Aspekte der Belebtheitsklassifikation abzielen. Zum einen eine Überprüfung der theoretischen Annahmen zu Belebtheit im Deutschen: wenn die grammatischen Phänomene, die im Allgemeinen mit Belebtheit in Verbindung gebracht werden, korrekt sind, sollte ein Klassifizierer, der diese Aspekte einbezieht, entsprechende Resultate liefern. Zum anderen werde ich einen lemmabasierten Ansatz verfolgen, der also nicht eine kontextbezogene Klassifizierung, sondern eine auf das Lexikon bezogene Unterteilung vornimmt. Ich werde mich stark an dem Ansatz aus Øvrelid (2006) orientieren, da er für eine Klassifikation, die auf Lemmata abzielt und weder manuelle Golddaten noch kontextuelle Belebtheitsklassifikation zur Verfügung hat, am geeignesten erscheint. Außerdem verspricht die vielfältige Verzahnung von Morphologie und Syntax im Deutschen ein gutes Ergebnis bei der Nutzung von morphosyntaktischen Merkmalen bei der Klassifikation. Da bislang keine automatische Belebtheitsklassifikation für das Deutsche vorgenommen wurde, soll diese Arbeit vor allem dazu dienen, die allgemeine Baseline festzulegen und die Komplexität des Problems zu umreißen. Dazu ist es hilfreich, sich mit den Methoden anderer Arbeiten zu anderen Sprachen auseinanderzusetzen und ihre Performanz für das Deutsche zu testen. Zum anderen sollte zunächst versucht werden, eine möglichst einfache Feature-Extraktion anzustreben und möglichst wenige Ressourcen zu verwenden, um das System bei eventueller schlechter Leistung in weiteren Studien zu erweitern.

In Abschnitt 2 werde ich auf die Definition von Belebtheit eingehen und inwieweit die Beschäftigung mit Belebtheit und der automatischen Erkennung von Belebtheitsmerkmalen hilfreich für weitere computerlinguistische

Gebiete und Aufgaben sein kann. Abschnitt 3 behandelt den derzeitigen Forschungsstand zu der Thematik und beleuchtet Methoden, die für andere Sprachen verwendet wurden, um eine Belebtheitsklassifikation vorzunehmen. Im Hauptteil, bestehend aus den Abschnitten 4 und 5, werde ich meine Methodik für das Deutsche und die verwendeten Daten vorstellen, sowie die Ergebnisse auswerten und diskutieren. Abschnitt 6 stellt ein abschließendes Résumé dar.

2 Belebtheit

Im Folgenden soll eine Übersicht über die unterschiedlichen Modellierungen von Belebtheit als sprachlicher Kategorie gegeben werden, sowie mögliche Anwendungsbereiche, in denen Wissen um Belebtheit für die maschinelle Sprachverarbeitung relevant sein kann.

2.1 Definition von Belebtheit als lexikalische Kategorie

Belebtheit ist eine lexikalische Kategorie, die eng mit dem Weltwissen und der kognitiven Verarbeitung der Welt durch die menschliche Psyche verbunden ist. Dies erklärt die allgemein angenommene Aufteilung von Belebtheitsbewertungen in der menschlichen Sprache in nicht nur BELEBT und UNBELEBT, sondern häufig auch in MENSCHLICH und NICHT-MENSCHLICH. Möchte man diese Kategorien nun hierarchisch absteigend nach ihrem Belebtheitsgrad anordnen, gelangt man zu folgender Darstellung:

(1) a. BELEBT > UNBELEBT

b. MENSCHLICH > NICHT-MENSCHLICH

(Yamamoto, 1999, S. 9).

Die exakte Hierachisierung der beiden Skalen in Bezug *aufeinander* ist indessen nicht ganz unproblematisch. Ist man zunächst geneigt, die Kategorie BELEBT durch die Anordnung aus (1b) zu ersetzen, und damit eine Hierarchisierung der Form

(2) MENSCHLICH > TIERISCH > UNBELEBT
(Silverstein, 1976)

zu erhalten, gibt es Ansätze, die tatsächlich nur eine Unterscheidung annehmen, wie sie in (1b) gezeigt ist und unter NICHT-MENSCHLICH sowohl tierische als auch unbelebte Entitäten fassen. In diesen Modellen wird der Kategorie MENSCHLICH also eine dominante Rolle in der Sprachverarbeitung zugesprochen. Auf der anderen Seite ist es nicht direkt eingängig, weshalb ein Mensch „belebter“ sein soll als ein beliebiges Tier (vgl. Yamamoto, 1999, S. 9ff.).

Es wurden zahlreiche Versuche gemacht, die von Silverstein (1976) vorgeschlagene Skala zu verfeinern und Annahmen zur Struktur menschlicher

Sprache mitzumodellieren. Foley und Van Valin (1985, S. 288) schlagen folgende Skala vor, die zugleich eine Skala der Salienz ist:

(3) Sprecher/Höhrer > 3. Person Pronomen > menschliche Eigennamen > menschliche „normale“ Nomen > andere belebte Nomen > unbelebte Nomen

Dazu kommen verschiedene Zweifelsfälle bezüglich der Belebtheit bestimmter Entitäten. Beispielsweise ist es nicht selten, dass eigentlich unbelebte Entitäten personalisiert werden, wie etwa Computer oder andere Arten von Maschinen, da sie etwas von selber zu tun scheinen: „Computers sometimes look as if they are thinking“ (Yamamoto, 1999, S. 18). Anders herum ist es ebenso möglich, dass Personen mit ihrem Amt oder ihrer Funktion angeredet und somit institutionalisiert werden und damit auf abstrakte Weise an Belebtheit verlieren können (vgl. Yamamoto, 1999, S. 4).

Belebtheit wird in den Sprachen der Welt durchaus unterschiedlich bewertet, da es sich zum einen um eine sprachliche Kategorie handelt, zum anderen aber auch um eine Kategorie, die stark mit dem Weltwissen verbunden ist und auf außersprachlichem Wissen basiert (vgl. Bloem & Bouma, 2013, S. 83).

Belebtheit wirkt sich beispielsweise auf morphologische Strukturen aus. Im Russischen werden maskuline belebte Nomen im Akkusativ systematisch anders dekliniert als maskuline unbelebte Nomen:

(4) a. нов-ый студент
now-yj student
neu-MASC.NOM.SG student.MASC.NOM.SG

'der neue Student'

b. нов-ого студент-а
now-ogo student-a
neu-MASC.ACC.GEN.SG student-MASC.ACC.GEN.SG

'den neuen Student'

(5) a. нов-ый диалог
now-yj dialog
neu-MASC.NOM.ACC.SG student.MASC.NOM.ACC.SG

'der neue Dialog' | 'den neuen Dialog'

Unbelebte Nomen sind im Akk. Sg. parallel zu ihrer Nominativ-Form endungslos. Belebte Nomen erhalten die Endung -a, identisch zu ihrer jeweiligen Genitiv-Form. Auch die Form нового ist eine genitivische Adjektiv-Form.

Das Japanische verwendet verschiedene Klassen an Numerus-Markern, die sich an Kategorien wie Belebtheit, Form oder Funktion der Entität orientieren (Baker & Brew, 2010, S. 54). Beispiele sind in Tabelle 1 (S. 9) zu finden.

Animacy		Shape		Function	
-nin	*hito-ga san-nin* 'three people'	*-hon*	*ohashi-ga san-bon* 'three chopsticks'	*-dai*	*kuruma-ga san-dai* 'three cars'
-hiki	*hebi-ga san-biki* 'three snakes'	*-mai*	*shatsu-ga san-mai* 'three shirts'	*-hatsu*	*juusei-ga san-batsu* 'three gunshots'
-tou	*uma-ga san-tou* 'three horses'				
-wa	*tori-ga san-wa* 'three birds'				

Tabelle 1: Japanische Numeral-Klassifikatoren (Baker & Brew, 2010, S. 54).

Im Deutschen werden die Fragepronomen *wer* und *was* danach ausgewählt, ob ihr Antezedens eine belebte oder unbelebte Entität darstellt:

(6) a. Hans hat gestern lange geschlafen.

b. Wer/*Was hat gestern lange geschlafen?

(7) a. Das Fussballspiel hat mir gut gefallen.

b. *Wer/Was hat dir gut gefallen?

Gleichzeitig wird durch die Auswahl des Pronomens vom Fragenden suggeriert, ob er nach einer belebten oder unbelebten Entität fragt, sollte der Antezedens nicht bekannt sein. Führt der Antezendens eine Tätigkeit aus, die nur von einer belebten Entität ausgeführt werden kann, muss das Pronomen *wer* gewählt werden, es sei denn, der Kontext lässt eigentlich unbelebte Entitäten als belebt zu (zum Beispiel in fiktiven Erzählungen):

(8) Wer/?Was hat mit mir geredet?

Weitere Beispiele für Pronomen mit Abhängigkeit zur Belebtheit des Antezedens sind *jemand*, *niemand* (belebter Antezedens), oder *etwas*, *nichts* (unbelebter Antezedens).

2.2 Verwendung von Belebtheitsmerkmalen in Natural Language Processing

Es stellt sich die Frage: Warum überhaupt eine Belebtheitsklassifikation durchführen? Die Antwort ist: Belebtheit spielt in vielen Sprachen eine Rolle bei der Verteilung grammatischer Rollen und bestimmter referentieller Auflösungsprozesse.

Kenntnisse über die Belebtheit von Wörtern kann unter anderem in folgenden wichtigen NLP-Tasks eine Rolle spielen: Anaphern- und Koreferenzresolution, Textgenerierung, maschinelle Übersetzung sowie automatisierte Diskurs-Modellierung.

Zaenen et al. (2004, S. 118) arbeiten die Bedeutung der verschiedenen möglichen grammatisch-diskursrelevanten Skalen (Genus, Person und Belebtheit) heraus und geben zu bedenken, dass alle diese Skalen sprachliche Einheiten beinflussen im Hinblick auf ihre Salienz im menschlichen Diskurs. Somit ist es nötig, Belebtheit zu erkennen, will man Diskurse automatisiert modellieren. Hier unterscheiden sich computerlinguistische und theoretisch-linguistische Ansätze, da letztere die Bedeutung solcher Skalen bereits lange endeckt haben und verwenden (Zaenen et al., 2004, S. 118). In der Textgenerierung ist die Bedeutung der Belebtheit relativ eindeutig. Es ist von großem Vorteil, bei Prädikaten zu wissen, welche Belebtheit den Rollen der Argumentstellen in der Regel zukommt. Beispielsweise wird im Allgemeinen ein Subjekt häufiger belebt sein, da es statistisch häufiger die semantische Rolle des Agens einnimmt. Agentitvität ist eine grammatisch-semantische Kategorie, die sehr stark mit Belebtheit zusammenhängt, da hier aktiv handelnde Entitäten der Verb-Handlung auftreten (vgl. Yamamoto, 1999, 148).

Es ist auch möglich, dass bestimmte Verben ausgehend von ihrer Semantik bestimmte Belebtheitsklassen für ihre subkategorisierten Phrasen benötigen:

(9) a. Der Mann/*?Baum bemerkte den kleinen Vogel.

b. Der Mann schaltete den Generator/*Hund ein.

Einschalten erfordert ein direktes unbelebtes Objekt. Die Semantik von *einschalten* lässt ein belebtes Objekt nicht zu, wie in (9b) zu sehen. (9a) zeigt

das Verb *bemerken*, das von einer belebten Entität als Subjekt ausgeht. Interessant ist hier, dass die Rolle von *der Mann* kein Agens ist, da Belebtheit nicht zwangsläufig immer mit der Agens-Rolle verknüpft sein muss. Zum anderen ist je nach Kontext auch ein belebter Baum denkbar (man denke hier vielleicht an Filme oder Märchen), weshalb ein relativierendes Grammatikalitätsurteil gesetzt ist. Hier wird deutlich, dass Belebtheit nur eine Tendenz darstellt, die kontextabhängig ist (in (9b) ist auch ein Roboter-Hund denkbar) und NLP-Anwendungen so immer davon profitieren werden, wenn Belebtheitsmerkmale zu Verfügung stehen. Auf diese Weise können dann Wahrscheinlichkeiten aufgestellt werden, wie sicher ein bestimmtes Verb mit einer bestimmten Selektion entsprechende Nomen akzeptiert und so ein komplexes, auf Diskurs und Kontext abzielendes System geschaffen werden.

Auch in der maschinellen Übersetzung ist es hilfreich, Belebtheit zu erkennen und um die verschiedenen Möglichkeiten der Sprachen zu wissen, Belebtheit auszudrücken. Je nach Sprache kann, wie in Abschnitt 2.1 gezeigt, Belebtheit unterschiedliche grammatische Konstruktionen bedingen. Diese Konstruktionen nicht zu erkennen und die diese Konstruktionen vorgebende Belebtheit nicht zu erkennen, muss zwangsläufig zu Schwierigkeiten in der automatischen Übersetzung führen (vgl. Zaenen et al., 2004, S. 119).

Øvrelid und Nivre (2007) konnten zeigen, dass Kenntnisse von (unter anderem) Belebtheitsklassifikationen die Parsing-Ergebnisse für Dependenzparser im Schwedischen signifikant verbessern können. Die Fehlerrate für Dependenzparsing auf einer großen Baumbank konnte hierbei um bis zu 50% gesenkt werden.

Im Deutschen ist Belebtheitserkennung relevant für die bereits vorgestellte Anaphernresolution bei Pronomen der Kategorie *wer* und *was*. Aber auch für eine sprachunabhängige Anaphern- und Koreferenzauflösung ist Belebtheit wichtig. Im Englischen werden Pronomen in der Regel nur nach Belebtheit ihrer Antezedenten ausgewählt, das Wissen um belebte Antezedenten-Kandidaten kann die Resolution somit extrem vereinfachen (siehe Evans & Orăsan, 2000). Aber auch im Deutschen, wo das Pronomen häufig nicht auf den Belebtheitsstatus des Antezedens schließen lässt, kann Belebtheit im Zusammenspiel mit anderen Faktoren einem Resolutions-Algorithmus durchaus wichtige Hinweise liefern. Generell sind für NLP im Deutschen aber auch alle bereits genannten Punkte relevant, sowie Aufgaben, die nur indirekt von Belebtheitserkennung profitieren.

3 Forschungsstand

Im Bereich der automatischen Belebtheitsklassifikation sind bislang relativ wenige NLP-Anwendungen entwickelt worden. Dies kann vor allem daran liegen, dass viele Ressourcen der maschinellen Sprachverarbeitung auf dem Englischen basieren, in dem Belebtheit für Grammatikalitätsurteile nur eine untergeordnete Rolle spielt. Entsprechend sind im Englischen nur wenige Annotationen von Belebtheit verfügbar, was sich auch auf die Annotationsschemata anderssprachiger Korpora auswirkt, die oftmals an englischen Vorbildern orientiert sind (vgl. zu der Problematik auch Bloem & Bouma, 2013, S. 82).

Nichtsdestotrotz sind Markierungen von Belebtheit für viele andere Sprachen relevant. Gerade im Bereich der Textgenerierung kann eine Klassifizierung von Belebtheit für die Wahl von Subjekt und Objekt hoch relevant sein, und zwar unabhängig von der verwendeten Sprache. Daneben reihen sich Anwendungsbereiche ein, die sprachspezifisch sind und vor allem auf Sprachen zutreffen, in denen Belebtheit ein wichtige Rolle spielt und fest in das grammatische System eingebettet ist. Die folgende Darstellung soll eine Übersicht zum Forschungsstand von Belebtheitsklassifizierung sowohl zum Englischen als auch zu weiteren Sprachen bieten.

3.1 Untersuchungen zum Englischen

Orăsan und Evans (2001) nutzen ihre Klassifikation zur Belebtheit englischer Nomen, um so leichter Anaphernresolution von Pronomen, die im Englischen Belebtheit anzeigen (*he*, *him*, *his*, *himself*, *she*, *her*, *hers*, *herself*) und dem potentiell belebten Antezendes zu betreiben. Ihre Methode, um die belebten Nomen zu ermitteln, beruht auf der Verwendung der semantischen Ressource *WordNet*[1], die unter anderem Synonyme, Hyponyme bzw. Hyperonyme für den englischen Wortschatz abdeckt. Die Idee ist, Bedeutungen von Terminalen in der Hyponym-Hyperonym-Hierarchie zu ermitteln und davon ausgehend die Belebtheit der übergeordneten Hyperonyme zu erschließen. Orăsan und Evans (2001) gehen davon aus, dass immer dann, wenn ein Knoten eines semantisch generelleren Wortes ausschließlich Knoten (Hyponyme) enthält, die belebt sind, auch das Hyperonym belebt sein muss. Um die Gefahr von Annotationsfehlern zu umgehen, nutzen sie das Chi-Quadrat-Maß (χ^2), falls ein Hyperonym sowohl belebte als auch unbelebte Hyponyme enthält, um das

[1] siehe Fellbaum (1998)

Ergebnis zu glätten. Eigennamen stellten eine große Herausforderung für den Klassifizierer dar, da sie meist nicht in WordNet enthalten sind und wurden entsprechend nicht berücksichtigt. Außerdem enthielten die Bedeutungen der Wörter keine Gewichtung, sodass eine ungleiche Verteilung von belebten Bedeutungen bei einem Lexem keine Auswirkung auf die Wahrscheinlichkeit hatte, im Kontext eher als belebt oder unbelebt klassifiziert zu werden.

In Orăsan und Evans (2007) wurde die Methode weiter verfeinert. Die Autoren stellen einen komplexeren Algorithmus vor, der anhand der gelernten Belebtheit für einzelne NPs operiert (Orăsan & Evans, 2007, S. 86). Die erreichte Accuracy zur Belebtheitsextraktion aus WordNet ist jedoch weitgehend identisch zu Orăsan und Evans (2001) und liegt bei durchschnittlich 97%.

Bowman und Chopra (2012) führten eine automatische Klassifikation von belebten NPs im Englischen durch und verwendeten dazu eine feiner spezifizierte Klassifikationsaufteilung als nur die Unterscheidung *belebt - unbelebt*. Das Korpus bestand ausschließlich aus gesprochenem Englisch. Ihr Klassifikations-Schema umfasst 10 Unterklassen, die die unterschiedlichen Grade an Belebtheit versucht zu erfassen (HUMAN, ORG(anizations), ANIMAL, MAC (automata), VEH(icles), PLACE, TIME, CONCRETE, NONCONC(rete) und MIX). Die Autoren stellen die These auf, dass neben Bedingungen innerhalb der Konstituenten auch das Prädikat des Satzes einen Einfluss auf die Klassifikation haben sollte, da bestimmte Verben nur spezifische Belebtheitskategorien an ihren subkategorisierten Phrasen erlauben. Interessanterweise führt eine Berücksichtigung von syntaktischen Merkmalen, die außerhalb der untersuchten NP liegen, jedoch nicht zu einer signifikanten Verbesserung der Ergebnisse. Als alleiniges Kriterium angewendet liefert Außer-Konstituenten-Wissen lediglich eine *Accuracy* von 50%, was auf Baseline-Niveau liegt und somit überraschend niedrig ist. Bei einer Anwendung aller Feature erreichen die Autoren eine Accuracy von etwa 85%.

3.2 Untersuchungen zum Norwegischen

Einen Ansatz unter Verwendung von Entscheidungsbäumen (Decision Trees) auf Grundlage von norwegischen automatisch getaggten Sprachdaten schlägt Øvrelid (2006) vor. Sie nimmt eine Klassifikation bezüglich der binären Entscheidung ANIMATE - INANIMATE vor und nutzt als Feature morphologische Kennzeichnungen, die im Norwegischen typisch für Belebtheit sind. Die Fea-

ture sind im Einzelnen: Häufigkeit von Subjekt- und Objektstellung in transitiven Verben, Agens-Rolle in Passivkonstruktionen, Anaphorische Beziehung von Personal- und Reflexivpronomen zum Antezedens sowie Possessivität durch den norwegischen Genitivmarker *-s*. Die Feature-Vektoren enthalten die relative Frequenz, mit der ein Nomen mit dem jeweiligen Merkmal auftritt, sowie die Form des Nomens und seine Klassifizierung (belebt, unbelebt). Es zeigte sich, dass für hochfrequente Nomen eine Accuracy von 87,5% erreicht werden konnte, wenn alle Feature verwendet wurden. Allerdings nimmt die Klassifikationsgenauigkeit dramatisch ab, sobald die Frequenz der Nomen sinkt. Durch eine Untersuchung zum Grad, mit dem einzelne Feature das Resultat verbessern, zeigte sich eine deutliche Wichtigkeit der Feature SUBJ(ekt), OBJ(ekt) und GEN(itiv), wobei OBJ wohl das stabilste und zuverlässigste Feature darstellt. Diese drei Feature scheinen demnach besonders vielversprechend zu sein, wenn es darum geht, im Norwegischen eine Belebtheitsklassifkation auf Grundlage von niedrigfrequenten Nomen bzw. einer geringen Datengrundlage vorzunehmen.

3.3 Untersuchungen zum Niederländischen

Bloem und Bouma (2013) schlagen eine dreigeteilte Belebtheitsklassifikation für Niederländisch vor, bestehend aus HUMAN, NONHUMAN und INANIMATE. Dabei kombinierten sie Golddaten aus einer lexikalischen Ressource mit Kontextdaten aus einer annotierten Baumbank, um so ein Trainingsset zu erhalten. Im Wissen, dass Belebtheit durchaus kontextabhängig ist, gingen sie in Ermangelung eines für Belebtheit annotierten Korpus davon aus, dass die Fälle, in denen Nomen ambig belebt oder unbelebt sein können, nicht allzu hoch sein dürften. Sie erreichen eine endgültige Accuracy von 92,5%, müssen jedoch einräumen, dass bei einer Baseline von 80% das Ergebnis besser sein könnte. Die Baseline basiert auf einer Klassifikation aller Nomen als INANIMATE. Vor allem die Klasse NONHUMAN stellte sich als problematisch heraus, da hier die Extraktionsvorgaben zu schwammig waren, sodass besonders hier Zweifelsfälle landeten, die biologisch nur schwach belebt sind und damit gleichzeitig sehr unwahrscheinlich als Agens in menschlicher Sprache fungieren. Auf einem balancierten Korpus war es möglich, alle drei Klassen zu unterscheiden, jedoch auf Kosten der Accuracy, die nun bei rund 72% lag (bei einer neuen Baseline von 40%).

In Karsdorp, van der Meulen, Meder und van den Bosch (2015) wird ein

überzeugendes Verfahren zur Belebtheitserkennung speziell in Erzählungen vorgestellt. Die Autoren ließen dazu 74 Volkserzählungen von Hand nach kontextueller Belebtheit annotieren, also mit Berücksichtigung der Belebtheit im Rahmen der Erzählung. Dadurch entstanden Ambiguitäten zur Belebtheit innerhalb desselben Lexems. Da sie nicht die Lemmata, sondern die Tokenebene untersuchten, verwendeten Karsdorp et al. (2015) ein n-Gramm-Modell, also eine Spanne von n Wörtern um einen Nukleus. Der Nukleus ist dabei das zu bestimmende Wort und die linken und rechten Kontexte werden durch $n = 3$ bestimmt. Außerdem verwendeten sie morphologische und syntaktisch-funktionale Feature. Desweiteren bestimmten sie ein semantisches Netz, um Ähnlichkeiten zwischen belebten Nomen abzubilden, mittels einer *Principle Component Analysis*. Bereits ein einfaches Verwenden des Trigramm-Klassifizieres erreicht einen F-Score von 0.98 für unbelebte und 0.85 für belebte Nomen. Das Hinzufügen der weiteren Feature bringt nur Verbesserung in Precision, Recall und F-Score, mit dem besten Ergebnis bei der Kombination *n-gramm - Wortart - Semantisches Netz* (F-Score: 0.99 für unbelebt, 0.93 für belebt). Im weiteren Vorgehen produzieren sie außerdem ein interessantes semantisches Cluster von belebten Wortgruppen und Nomen aus einem „übernatürlichen“ Kontext, die die Volkssagen bevölkern.

3.4 Untersuchungen zum Japanischen

Baker und Brew (2010) führten eine Untersuchung zur Belebtheitsklassifikation im Japanischen durch, die sie durch Hinzunahme von englischen Daten erweiterten. Dazu zählten sie Frequenzen von Nomen, die als Subjekt, bzw. Objekt eines transitiven Verbs auftraten. In einem weiteren Experiment ersetzten sie die einfache Frequenz durch einen Wert, der sich aus der Subjektfrequenz belebter Nomen jedes ermittelten Verbs geteilt durch die Gesamtzahl der Subjekte ergab. Dadurch wird ein Fokus auf die Eigenart eines Verbs gelegt, belebte Subjekte zu bevorzugen, oder eben nicht. Da die verbleibenden Verben nur 40% des Vorkommens abdeckten, bildeteten sie Suffix-Klassen, um so Felder von Nomen zu erzeugen, die sie als Klassen vor der Klassifikation verwendeten. So konnten auch Nomen erfasst werden, die im Training nicht gezählt werden. In einem dritten Experiment nutzten sie englische Sprachdaten und bestimmten die Belebtheit der englischen Lexeme. Der Grund dafür war die große Anzahl an englischen Lehnwörtern in den japanischen Daten. Davon ausgehend, dass die lexikalische Belebtheit der englischen Lehnwörtern den englischen Originalen enspricht, konnte so schlussendlich eine Daten-

spanne von 97% abgedeckt werden, mit einer Accuracy von gemittelt 86%.

3.5 Übergreifende Ergebnisse

In allen Arbeiten zeigte sich, dass sich die Klassifikation von unbelebten Nomen deutlich einfacher gestaltet als die von belebten Nomen. Grund dafür ist das offenbar sprachübergreifende Phänomen, dass belebte Nomen nur einen kleinen Teil der Nomen ausmachen, der wesentlich größere Anteil wird von unbelebten Nomen bestimmt. Somit tendiert ein Klassifizierer auf einer entsprechend großen Textmenge immer dazu, im Zweifelsfall das Nomen als unbelebt zu klassifizieren.

Weiterhin scheinen ebenfalls sprach- und untersuchungsübergreifend besonders syntaktische Funktionen, hier besonders Subjekt- und Objektpositionen bei transitiven Verben eine besonders hohe Accuracy zu erzielen. Diese Funktionen scheinen sich demnach besonders anzubieten, um Feature für eine Belebtheitsklassifikation aufzustellen.

4 Ein Belebtheitsklassifizierer für Deutsch

Im Folgenden werde ich die von mir verwendeten Daten und Methoden vorstellen, die ich genutzt habe, um eine automatische Belebtheitsklassifikation für das Deutsche zu realisieren.

4.1 Die Feature

Es liegt die Vermutung nahe, dass Nomen in Subjekts-Position statistisch wesentlich häufiger belebt sind als andere Nomen. Da das Subjekt oftmals die semantische Rolle des Agens erhält, ist gerade diese Funktion oftmals Träger belebter Information. Vor allem in transitiven Verbrelationen sollte dies deutlich werden. Das Subjekt sollte statistisch häufiger die semantische Rolle des Agens, das direkte Objekte häufiger die Rolle des Patiens belegen. Zwei der verwendeten Feature sind deswegen SUBJ und OBJ, die jeweils eine relative Frequenz angeben: die Anzahl des Auftretens als Subjekt eines transitiven Verbs gemessen am Gesamtvorkommen als Subjekt. Bezogen wird diese Frequenz also jeweils auf ein Lemma. Berücksichtigt werden nur die nominalen Köpfe eines Subjekts, also keine tiefer eingebetteten Phrasen wie z.B. PPs. Für das Feature OBJ gilt analoges im Hinblick auf direkte, also Akkusativ-Objekte.

Auch Nominalphrasen im Vorfeld sollten häufig belebt sein, da das Vorfeld im Deutschen mit dem sogenannten *Topic* und *Focus* verknüpft ist (vgl. Berman, 2000, S. 25). Topic ist ein Begriff der Diskurstheorie, der neu eingeführte Diskursreferenten erfassen soll und in dieser Definition soll Topic im Weiteren behandelt werden (zu beachten ist beim Begriff Topic die große Begriffsverwirrung und oftmals schwammige Definition in der linguistischen Literatur (vgl. auch Yamamoto, 1999, S. 60)). Da Topic häufig mit Subjekt-Vorkommen zusammenfällt und sich nach obiger Definition auch aus diskurstheoretischer Sicht für belebte Entitäten anbietet, ist die Verbindung von Topic und Belebtheit tatsächlich gegeben (Yamamoto, 1999, S. 60–67). Zwar ist das Subjekt im Vorfeld nicht zwangsläufig als Topic oder Focus gekennzeichnet (Berman, 2000, S. 25), aber auf großen Datenmengen sollte trotzdem eine Tendenz erkennbar sein, dass Subjekte, die im Vorfeld stehen, gute Kandidaten für Belebtheit sind, als auch Nicht-Subjekt-Phrasen, die dennoch in Topic- oder Focus-Position stehen. Steht das Subjekt nicht im Vorfeld, sondern im Mittelfeld, kann in das Vorfeld auch eine andere Konstituente als das Subjekt treten. Handelt es sich hierbei um eine NP, z.B. um ein in das Vor-

feld bewegtes Objekt, kann die These aufgestellt werden, dass dieses Objekt tendenziell eher belebt sein könnte, da es im Kontext als Topic, bzw. Focus gekennzeichnet wird. Da eine NP im Vorfeld jedoch in der Regel das Subjekt sein wird und andere eingebettete NPs im Vorfeld, z.B. innerhalb einer PP, nicht berücksichtigt werden können und sollen, bietet sich hier die Verwendung einer absoluten Frequenz pro Lemma an. Für jedes Nomen wird also gezählt, wie oft es im Vorfeld steht. Ist das Nomen auch gleichzeitig Kopf einer Subjekts-NP, sollten sich diese Informationen addieren und die Belebtheit noch wahrscheinlicher werden lassen. Das entsprechende Feature für die Vorfeldposition heisst VF.

Zum Dritten lässt sich für das Deutsche eine morphologische Markierung festmachen, die stark auf Belebtheit hindeutet: die *schwachen Maskulina* (im Folgenden sw. Mask.). Sw. Mask. kennzeichnen sich durch eine im deutschen Flexionssystem einmalige Deklination innerhalb der Gruppe der maskulinen Nomen aus. Im Gen., Akk. sowie Dat. Sg. enden diese Nomen auf *-en* oder auf *-n*, falls der Stamm auf *-e* endet. Maskuline Nomen der starken, bzw. gemischten Deklination enden im Gen. Sg auf *-s* und sind im Akk. und Dat. Sg. endungslos (vgl. Eisenberg, 2013, S. 153-155). Die Nomen, die dieser Deklinationsklasse angehören, sind in der Regel belebt und oftmals mit der Klasse MENSCH zu versehen (Eisenberg (2013, S. 154) und Köpcke (2005, S. 71)). Köpcke (2005) führt ein sehr differenziertes System zur Kategorie der sw. Mask. vor, in dem er semantische Merkmale wie das der Belebtheit und prosodische Merkmale wie auslautendes Schwa zu einer Skala zur Erkennung von sw. Mask. zusammenführt. Obwohl Köpcke (2005, S. 71) zu bedenken gibt, dass vor allem das auslautende Schwa sw. Mask. kennzeichnet, geht aus seiner Prototypikalitätsskala der sw. Mask. doch hervor, dass ein Nomen umso belebter ist, je wahrscheinlicher es als sw. Mask. dekliniert wird. Anhand der Flexionsendung der maskulinen Nomen sollten sich also bei entsprechender Datenmenge die sw. Mask. ableiten lassen. Obwohl maskuline Nomen nur einen Teil der deutschen Nomen ausmacht und innerhalb der maskulinen Nomen auch eher die starke Flexion dominiert anstatt die schwache (Köpcke, 2005, S. 71), kann dieses Feature dennoch nützlich sein, um Zweifelsfälle bei der Klassifikation zu beseitigen und dem Klassfizierer eine starke regelbasierte Richtlinie liefern. Vor der Belebtheitsklassifikation muss also auch eine Sw.-Mask.-Klassifikation erfolgen, deren Entscheidungsbaum in Abbildung 1, S. 19 zu sehen ist.

Für jedes maskuline Nomen wird der Kasus und der Numerus überprüft. Handelt es sich um Akk., Dat. oder Gen. Sg. und das Nomen endet auf *-en*,

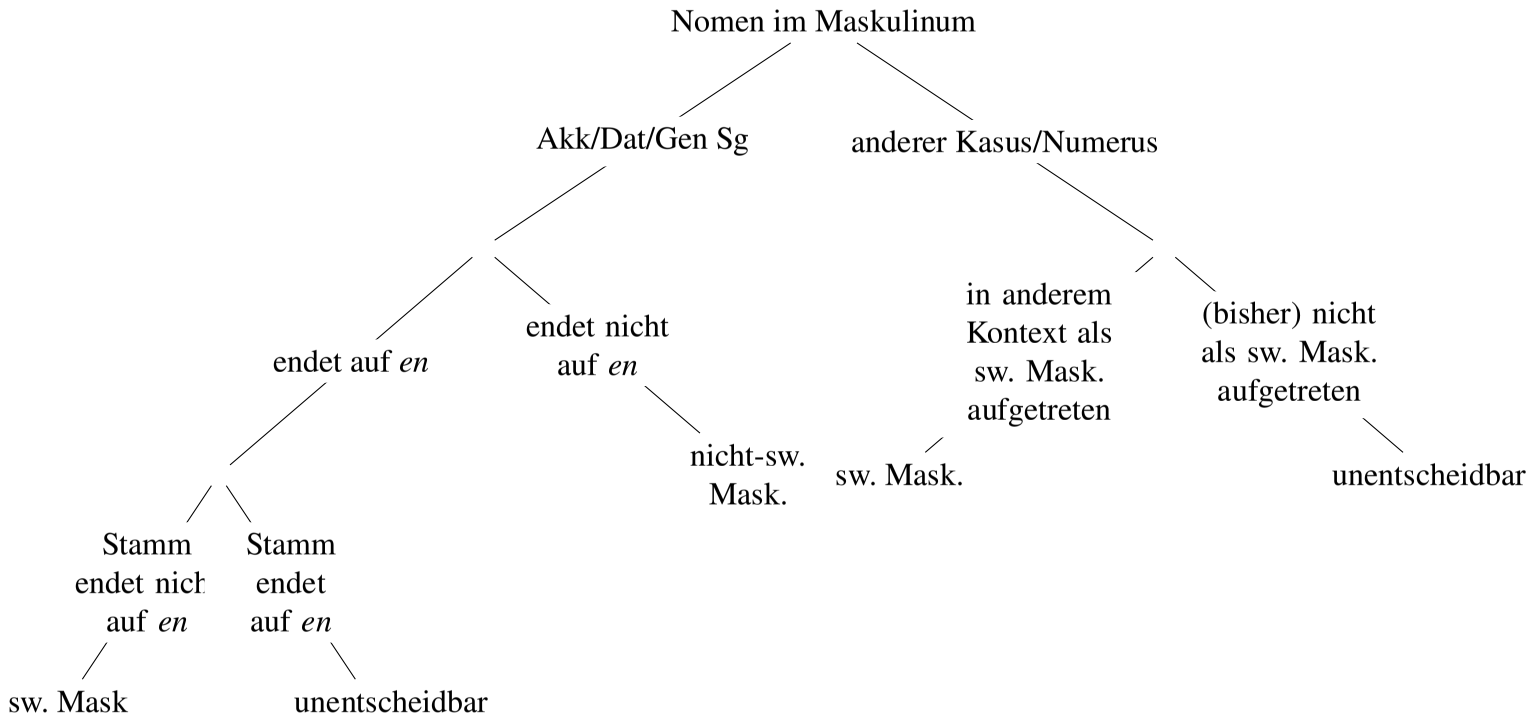

Abbildung 1: Entscheidungsbaum für die automatische Erkennung schwacher Maskulina

während der Stamm nicht auf *-en* endet, kann sicher von einem sw. Mask. ausgegangen werden. Endet der Stamm auf *-en*, muss die Entscheidung unentschieden bleiben, da dann nicht Fälle der Form *der Garten - dem Garten - den Garten* ausgeschlossen werden können, da es sich hier um ein starkes Maskulinum handelt, das im Dat. und Akk. Sg. endungslos bleibt (zu sehen am Gen. Sg. *des Gartens* anstatt **des Garten*). Endet der Akk./Dat./Gen. Sg. nicht auf *-en*, ist sicher von einem starken oder gemischten Mask. auszugehen. Für eine tokenbasierte Klassifikation müsste anschliessend zusätzlich geprüft werden, ob ein bis dahin als sw. Mask. klassifiziertes Nomen auch in Fällen auftritt, in denen es in einem anderen Kasus oder Numerus (d.h. Nominativ oder Plural) steht. So wäre eine kontextbasierte Klassifikation möglich. Da der in dieser Arbeit vorgestellte Ansatz jedoch ein lemmabasierter ist, kann hier auf diese zusätzliche Entscheidung verzichtet werden und ist auch nicht implementiert, sondern lediglich aus Gründen der Vollständigkeit in Abbildung 1 enthalten. Alle Nomen, die als feminin oder neutrum gekennzeichnet sind, werden als nicht-SWMASK klassifiziert.

Bei der Klassifizierung der Belebtheit finden also insgesamt folgende Feature Verwendung:

Feature	**Mögliche Werte**
SUBJ	relative Frequenz
OBJ	relative Frequenz
VF	relative Frequenz
SWMASK	ja, nein

Tabelle 2: Verwendete Feature für die Klassifizierung

4.2 Die Daten

Als Trainings- und Evaluations-Datenset verwende ich die umfangreiche Baumbank *TüBa-D/Z* der Universität Tübingen[2]. Die verwendete Versionsnummer lautet 6.0 (Telljohann, Hinrichs, Kübler, Zinsmeister & Beck, 2012), die erweitert zu den vorherigen Versionen eine Lemma-Annotation enthält. Das verwendete Export-Format ist das Negra-Export-Format in TigerXML[3] dargestellt.

TüBa-D/Z 6.0 besteht aus 976.262 Token, 1.165.657 Phrasen und 55.814 Sätzen. Jedem Token sind morphologische Eigenschaften wie Wortart (nach

[2] siehe http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html

[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html

den Guidelines des STTS[4]) beigefügt. Außerdem enthält das Negra-Export-Format 4 eine Lemma-Annotation. Über Index-Pointer werden die Phrasen markiert (siehe Abbildung 2, S. 21).

```
<s id="s1">
<graph root="s1_506">
<terminals>
 <t id="s1_1" word="Veruntreute" lemma="veruntreuen" pos="VVFIN" morph="3sit" />
 <t id="s1_2" word="die" lemma="die" pos="ART" morph="nsf" />
 <t id="s1_3" word="AWO" lemma="AWO" pos="NN" morph="nsf" />
 <t id="s1_4" word="Spendengeld" lemma="Spendengeld" pos="NN" morph="asn" />
 <t id="s1_5" word="?" lemma="?" pos="\." morph="--" />
</terminals>
<nonterminals>
 <nt id="s1_500" cat="VXFIN">
  <edge label="HD" idref="s1_1" />
 </nt>
 <nt id="s1_501" cat="EN-ADD">
  <edge label="-" idref="s1_3" />
 </nt>
 <nt id="s1_502" cat="NX">
  <edge label="HD" idref="s1_4" />
 </nt>
 <nt id="s1_503" cat="LK">
  <edge label="HD" idref="s1_500" />
 </nt>
 <nt id="s1_504" cat="NX">
  <edge label="-" idref="s1_2" />
  <edge label="HD" idref="s1_501" />
 </nt>
 <nt id="s1_505" cat="MF">
  <edge label="OA" idref="s1_502" />
  <edge label="ON" idref="s1_504" />
 </nt>
 <nt id="s1_506" cat="SIMPX">
  <edge label="-" idref="s1_503" />
  <edge label="-" idref="s1_505" />
 </nt>
</nonterminals>
</graph>
</s>
```

Abbildung 2: Der erste Satz der TüBa-D/Z 6.0 in TigerXML

Als Golddatengrundlage habe ich mich für das Synset-Lexikon *GermaNet*[5] der Universität Tübingen entschieden, dass bereits in Teilen in die TüBa-D/Z-Annotationen integriert ist (ab Version 9.1). GermaNet kann als deutsche Variante zum englischen *WordNet*[6] gesehen werden.

GermaNet 5.2 besteht aus insgesammt 84.859 Lemma-Einträgen für Adjektive, Nomen und Verben, davon 64.315 Nomen. Die nominalen Kategorien MENSCH und TIER können als Grundlage für einen Goldstandard an belebten Nomen genutzt werden und umfassen 10.392 Lemmata für MENSCH und 2410 für TIER. Damit beläuft sich die verwendete Datengrundlage für belebte Nomen auf insgesamt 12.802 Lemmata. Abbildung 3 (S. 22) zeigt einen Ausschnitt aus der Hyponomie-Relation zum Oberbegriff MENSCH. Die Hy-

[4] Stuttgart-Tübingen-TagSet (Schiller, Teufel, Stöckert & Thielen, 1999)
[5] siehe Hamp und Feldweg (1997) und Henrich und Hinrichs (2010)
[6] Fellbaum (1998)

ponyme sind in Synsets angeordnet. Damit ergibt sich eine überkreuzende semantische Zuordnung. Auf vertikaler Ebene bestehen Hyponomie-Hyperonomie-Relationen, auf horizontaler Ebene werden synonyme Beziehungen angezeigt.

```
<synset id="s37297" category="nomen">
<lexUnit id="l53687" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
<orthForm>Fußgänger</orthForm>
</lexUnit>
<lexUnit id="l53688" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
<orthForm>Fußgängerin</orthForm>
</lexUnit>
<lexUnit id="l53689" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
<orthForm>Passant</orthForm>
</lexUnit>
<lexUnit id="l53690" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
<orthForm>Passantin</orthForm>
</lexUnit>
</synset>
<synset id="s37298" category="nomen">
<lexUnit id="l53691" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
<orthForm>Fahrer</orthForm>
</lexUnit>
<lexUnit id="l53692" sense="1" source="core" namedEntity="no" artificial="no" styleMarking="no">
<orthForm>Fahrerin</orthForm>
</lexUnit>
</synset>
```

Abbildung 3: Ausschnitt der GermaNet-XML-Datei zur Kategorie *Mensch*

4.3 Die Tools

Die verwendete Programmiersprache, um die Daten zu extrahieren, ist die objektorientierte Programmiersprache Python[7] (in Version 3, auch kurz Python 3). In den 1990er Jahren von Guido van Rossum entwickelt, ist Python eine Sprache, die viele Vorteile bietet, wenn es um die Verarbeitung natürlicher Sprache geht. Vor allem zahlreiche Module lassen die Funktionsweise stark erweitern und bieten dem Programmierer eine große Auswahl an fertigen Funktionen, die in das eigene Projekt integriert werden können. Das *Natural Language Toolkit* (Bird, Loper & Klein, 2009) ist eine umfangreiche Python-Bibliothek, die sowohl mit eigenen Korpus-Daten als auch zahlreichen Werkzeugen zur Verarbeitung natürlicher Sprache aufwartet. Dies umfasst beispielsweise verschiedene Parser, Tools für die Frequenzanalyse, Klassifizierer und Korpusdaten sowie Tools zur Verarbeitung von Korpora.

Python bietet auch eine recht einfache Verarbeitung von Strings und Textdaten und stellt eine umfangreiche Unterstützung für reguläre Ausdrücke bereit, was die Sprache noch einmal prädestiniert für die Verarbeitung natürlicher Sprache macht.

[7] https://www.python.org/doc/

Als Tool zum überwachten maschinellen Lernen verwende ich das Open-Source-Tool C5.0[8], dass auch in Øvrelid (2006) zur Anwendung kommt. Mit diesem Tool ist es möglich, eingegebene Daten und Klassifikationen als Entscheidungsbaum zu realisieren. Darüber hinaus bietet es integrierte Modi zur automatischen Cross-Validierung und Boosting. Das Tool ist in C programmiert und bietet neben Entscheidungsbäumen auch daraus abgeleitete Regeln. C5.0 ist für Unix- und Windows-Systeme (für Windows *See5*) verfügbar.

C5.0 nimmt Entscheidungsprozesse anhand einer Vektorenmenge vor, die die einzelnen Datensätze repräsentiert. Dabei können den Feature verschiedene Werte-Klassen zugewiesen werden:

- **continous**: für Integer- und Float-Werte
- **f, t**: für binäre ja-nein-Entscheidungen
- **label**: zur Kennzeichnung und Zuordnung der Daten

Feature, denen nicht zuverlässig ein Wert zugewiesen werden kann, werden mit *?* markiert. Diese Feature gehen dann nicht in die Bewertung ein.

Beispielhafte Vektoren, wie sie für die Belebtheitsklassifizierung verwendet werden, sehen folgendermaßen aus:

```
jude,animate,0.13333333333333333,0.044444444444444446,0.111111111111111,t
beschäftigungskrise,inanimate,0,0,0,f
bettdecke,inanimate,0,0.5,0,f
vizeminister,animate,0,0,0,?
uni,inanimate,0.07575757575757576,0.045454545454545456,0.07575757575757576,f
usw.
```

Dabei sind die Zeilen wie folgt zu lesen:

- Lemma,Belebtheitsklasse,SUBJ,OBJ,VF,SWMASK

Die Werte für SUBJ, OBJ und VF können zwischen 0 und 1 liegen (0: trat nie als Subjekt etc. auf; 1: trat nur als Subjekt etc. auf). Die Werte für SWMASK sind entweder t(rue),f(alse) oder ? (konnte nicht bestimmt werden).

[8] https://www.rulequest.com/see5-unix.html

5 Evaluation

5.1 Verwendete Evaluations-Maße

Im Weiteren werden folgende Evaluationsmaße verwendet: *Precision*, *Recall*, *Accuracy*, *F-Score* (nach Manning & Schütze, 1999, S. 267ff.) und *Matthews Correlation Coefficient* (vereinfacht dargestellt nach Matthews (1975, S. 445)). Die Maße basieren auf den Daten der Konfusionsmatrizen[9], die bei einer Klassifizierung entstehen. Folgende Klassifizierungsfälle sind möglich:

- *True Positives (tp)*: Korrekte Klassifizierung der betrachteten Klasse
- *True Negatives (tn)*: Korrekte Klassifizierung der nicht betrachteten Klasse
- *False Positives (fp)*: Falsche Klassifizierung der betrachteten Klasse
- *False Negatives (fn)*: Falsche Klassifizierung der nicht betrachteten Klasse

Die betrachtete Klasse kann in dieser Untersuchung entweder BELEBT oder UNBELEBT sein. Je nach aktuell betrachteter Klasse sind verschiedene Werte für Precision (Maß der Trefferquote), Recall (Maß der Vollständigkeit) und F-Score (Harmonisches Mittel aus Precision und Recall) möglich:

- $P = \frac{tp}{tp+fp}$
- $R = \frac{tp}{tp+fn}$
- $F_\alpha - Score = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}}$ mit α als Gewichtung

In der Auswertung wird stets $\alpha = 1$ gesetzt, womit sich die F-Score-Formel wie folgt vereinfacht:

- $F_1 - Score = \frac{2\times P\times R}{P+R}$

Die Accuracy ist für alle Klassen immer identisch und stellt ein Maß für die korrekten und falschen Klassifizierungen in einem dar:

- $Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$

[9] siehe den Anhang ab S. 40 für sämtliche bei den Experimenten entstandenen Konfusionsmatrizen

Auch der Matthews Correlation Coefficient ist für alle Klasse identisch. Er kann Werte zwischen -1 und 1 annehmen. -1 bedeutet eine völlige Fehlklassifikation, 1 eine vollkommen korrekte Klassifikation und 0 eine Zufallsentscheidung. Der Koeffizient ist ideal geeignet, um binäre Entscheidung mit ungleich verteilten Klassen zu klassifizieren und damit das ideale Evaluationsmaß für das vorliegende Problem. Der MCC berechnet sich nach folgender Formel:

- $MCC = \frac{(tp \times tn) - (fp \times fn)}{((tp+fp)(tp+fn)(tn+fp)(tn+fn))^{\frac{1}{2}}}$

5.2 Auswertung der Daten

5.2.1 Erkennung der Belebtheit durch GermaNet-Synsets

Wie bereits in Abschnitt 4 beschrieben, verwende ich für die Golddaten-Erzeugung für Belebtheits-Annotationen die semantische Ressource GermaNet. Zwei Kategorien, die GermaNet anbietet, scheinen besonders vielversprechend für belebte Nomen zu sein: MENSCH und TIER. Diese zwei Mengen an Nomen setze ich als Quelle für belebte Nomen an. Danach extrahiere ich alle Nomen aus TüBa-D/Z und prüfe, ob sie in der Menge der belebten Nomen aus GermaNet enthalten sind. Um Komposita abzufangen, wie z.B. *IBM-Berater*, die nicht in GermaNet verzeichnet sind (das GermaNet-Lemma lautet lediglich *Berater*, bzw. *Beraterin*), prüfe ich jeweils, ob das Nomen aus GermaNet als rechter Rand des TüBa-D/Z-Lemmas auftritt. Ein einfacher Test, ob das GermaNet-Lemma generell im TüBa-D/Z-Lemma enthalten ist, wäre nicht sinnvoll, da deutsche Komposita in der Regel eine Rechtsköpfigkeit aufweisen und so Fälle wie *Berater-Ausweis* als belebt gekennzeichnet würden. Ich gehe hierbei davon aus, dass ein belebter Kopf seine Belebtheit an das Gesamtkompositum weitervererbt. Was durch eine Klassifikation mit GermaNet-Lemmata nicht abgefangen werden kann, sind Eigennamen, die ich entsprechend außen vor lasse. Die extrahierten Nomen sind also sämtlich mit dem STTS-Tag *NN* (*Normal Noun*) gekennzeichnet. Allerdings bietet die Art der Quelle für die Feature-Gewinnung eine Möglichkeit, Eigennamen zumindest indirekt abzufangen: da es sich bei TüBa-D/Z um ein Zeitungskorpus handelt, werden Eigennamen häufig die Anreden *Mann* und *Frau* vorangestellt. Diese Anreden sind als *NN* annotiert und werden bei der Extraktion entsprechend als belebte Nomen klassifiziert.

Für TüBa-D/Z 6.0 ergeben sich folgende Frequenzen an belebten und unbelebten Nomen:

Anz. Nomen (Lemma)	Anz. belebte Nomen (Lemma)	Anz. unbelebte Nomen (Lemma)	Anteil belebte Nomen
42666	8513	34153	19,95%

Tabelle 4: Verteilung der Belebtheit in TüBa-D/Z 6.0

Gut 20% der im Korpus auftretenen Nomen sind demnach belebt, 80% der Nomen sind unbelebt. Die Baseline für das Problem der Belebtheitsklassifikation liegt damit bei den vorliegenden Daten bei 80%. Das bedeutet, dass bei einer Klassifikation aller Datensätze mit dem Attribut UNBELEBT eine Fehlerrate von 20% besteht, also 20% der Daten tatsächlich eigentlich BELEBT sind. Für einen erfolgreichen Klassifizierer sollte die Accuracy also deutlich über 80% liegen, zumal die Baseline ohnehin sehr hoch ist.

5.2.2 Experiment 1: Binäre Feature

In einem ersten einfachen Experiment habe ich nur binäre Werte der Feature zugelassen. Trat ein Wort als Subjekt auf, wurde dem entsprechenden Lemma das Feature SUBJ = t zugewiesen. Trat ein Wort nie als Subjekt auf, bekam es das Feature SUBJ = f. Entsprechendes erfolgte für die restlichen Feature. Die Feature enthielten also nur Werte im Sinne *ja - nein*.

Feature	Mögliche Werte
SUBJ	ja, nein
OBJ	ja, nein
VF	ja, nein
SWMASK	ja, nein

Tabelle 5: Verwendete Feature für Experiment 1

Der Klassifizierer lief dabei auf einer Vektormenge von 34.134 Lemmata im Trainingsset und 8532 Lemmata im Testset. Dies entspricht einer Datenaufteilung von 80% für das Trainingsset und 20% für das Testset, was bei den übrigen Experimenten (abgesehen von Experiment 3) beibehalten wurde. Als Ergebnis klassifizierte das Programm fast alle Vorkommen als UNBELEBT. Nur das Feature SWMASK konnte als hilfreich für eine Klassifikation verwendet werden.

Folgende Klassifizierung hat sich für dieses Experiment ergeben. Die Werte (wie auch die im Weiteren folgenden Werte) wurden jeweils auf die zweite Nachkommastelle gerundet:

		BELEBT			UNBELEBT		
Accuracy	**MCC**	Precision	Recall	F-Score	Precision	Recall	F-Score
0.81	0.16	0.04	0.81	0.09	0.99	0.81	0.89

Tabelle 6: Ergebnisse Klassifizierung Experiment 1

Bei einer 10-fachen Crossvalidierung mit den gleichen Daten ergibt sich zwar die gleiche Accuracy, allerdings benutzt das Tool *C5.0* für die Crossvalidierung parallel auch regelbasierte Klassifizierungen, die bei 6 von 10 Datenaufteilungen folgenden Zusammenhang zwischen dem Feature SWMASK und der Klassifizierung herstellen:

```
Rule 1:
    swmask = t
    -> class animate [0.723]

Rule 2:
    swmask = f
    -> class inanimate [0.869]
```

Das bedeutet, dass ein Nomen mit dem Feature SWMASK $= t$ (es handelt sich also um ein schwaches Maskulin) mit etwa 70-prozentiger Wahrscheinlichkeit als belebt klassifiziert werden kann. Ein nicht-schwaches Maskulin kann mit 86%-iger Wahrscheinlichkeit als unbelebt klassifiziert werden.

Natürlich machen Maskulina und erst recht schwache Maskulina nur einen geringen Anteil der Gesamtdaten aus und haben somit kaum Einfluss auf die Gesamtklassifikation. Aber dennoch erkennt der Klassifizierer den in der theoretischen Forschung beschriebenen Zusammenhang zwischen schwacher Flexion bei nominalen Maskulina und der Belebtheit. Eventuell kann der Klassifizierer diese Zusatzinformation bei einer Featuregrundlage, die mehr Daten abdeckt, entsprechend einbringen.

5.2.3 Experiment 2: Gewichtung des SUBJ- und OBJ-Features

In einem weiteren Experiment wurden die Feature SUBJ, OBJ und VF neu gewichtet. Anstatt einer binären ja-nein Entscheidung werden jetzt relative Frequenzen (wie in Abschnitt 4.1 beschrieben) berechnet und jedem Lemma zugeordnet. Das Ergebnis der Klassifizierung ist in Tabelle 7 (S. 28) dargestellt.

Es zeigt sich, dass eine Belebtheitsklassifizierung anhand der gewählten Feature nicht sonderlich erfolgreich ist. Zwar erhält man eine Accuracy von

		BELEBT			UNBELEBT		
Accuracy	**MCC**	Precision	Recall	F-Score	Precision	Recall	F-Score
0.8	0.15	0.06	0.65	0.11	0.99	0.8	0.89

Tabelle 7: Ergebnisse Klassifizierung Experiment 2 mit sämtlichen Subjekten und Objekten

80%, jedoch relativiert sich dieser Wert, wenn, wie im vorherigen Abschnitt vorgeschlagen, eine Baseline von 80% anstatt 50% gelten soll. Hier ist es angebracht, auf Precision und Recall, bzw. F-Score zu schauen. Wie vermutet, neigt der Klassifizierer stark dazu, die mengenmäßig deutlich größere Klasse UNBELEBT sehr stark zu gewichten und sehr häufig zu wählen. Entsprechend liegt der F-Score für die Klasse UNBELEBT bei respektablen 0.89. Für eine durchgehend korrekte Klassifizierung von BELEBT scheinen die Feature nicht ausreichend aussagekräftig zu sein. Dies spiegelt sich in einem F-Score von 0.11 wieder. Wie bereits erläutert, liefert der MCC einen aussagekräftigeren Wert für die vorliegende Binär-Klassifikation als die Accuracy und ist mit 0.15 nah an der Zufallsentscheidung.

Die Ergebnisse einer zusätzlich durchgeführten 10-Cross-Validierung sind mit den vorliegenden Maßen nahezu identisch, der MCC liegt mit 0.16 jedoch geringfügig höher. Es zeigt sich, dass die Hinzunahme von relativen Frequenzen das Ergebnis nicht beeinflusst. Wo in Experiment 1 nur das Feature SWMASK verwendet wurde, konnte ein identischer MCC erreicht werden. Allerdings wurden die übrigen Feature nun in den Entscheidungsprozess einbezogen und ein entsprechender Entscheidungsbaum erstellt. Wie in Abbildung 4 (S. 29) zu sehen, erkennt der Klassifizierer durchaus eine Tendenz dazu, dass Nomen, die häufig als Subjekt auftreten, eher BELEBT sind, während Nomen mit einer niedrigen Subjekt-Frequenz eher als UNBELEBT klassifiziert werden. Das OBJ-Feature zeigt die entsprechend gegenläufige Tendenz. Als Hauptklassifizierung fungiert immer noch das Feature SWMASK. Für das VF zeigt sich die zur Erwartung entgegengesetzte Beobachtung. Bei einer höheren Frequenz von Nomen im Vorfeld tendiert der Klassifizierer dazu, die Nomen als UNBELEBT zu klassifizieren. Dies scheint der These entgegenzustehen, dass Topic und Belebtheit im Deutschen zusammenhängen. Eventuell ist das bloße Auftreten im Vorfeld jedoch auch nicht signifikant genug, um Topic im Deutschen abzubilden und das Feature VF deswegen schlecht gewählt.

Als nächstes wurden die Feature SUBJ und OBJ modifiziert, sodass nur noch

```
swmask = t: animate
swmask = f:
:...subj <= 0.1716418: inanimate
    subj > 0.1716418:
    :...obj > 0.1791045: inanimate
        obj <= 0.1791045:
        :...obj <= 0.005524862:
            :...subj > 0.4736842: inanimate
            :   subj <= 0.4736842:
            :   :...subj > 0.3581395: animate
            :       subj <= 0.3581395:
            :       :...vf <= 0.3529412: animate
            :           vf > 0.3529412: inanimate
            obj > 0.005524862:
            :...vf > 0.3157895:
                :...subj <= 0.2857143: inanimate
                :   subj > 0.2857143:
                :   :...obj <= 0.06896552: animate
                :       obj > 0.06896552: inanimate
                vf <= 0.3157895:
                :...subj > 0.375: animate
                    subj <= 0.375:
                    :...obj <= 0.0775862: animate
                        obj > 0.0775862:
                        :...vf <= 0.2380952: animate
                            vf > 0.2380952: inanimate
```

Abbildung 4: Entscheidungsbaum aus C5.0 für Experiment 2 mit relativen Frequenzen als Featurewerte

Subjekte und Objekte eines transitiven Verbes berücksichtigt wurden. Dies hat den Vorteil, dass so stärker auf die Agens-Patiens-Verteilung eingegangen werden kann. In transitiven Verbrelationen sollte statistisch häufiger ein belebtes Subjekt mit einem unbelebten Objekt auftreten. Diese Erwartung wurde nicht erfüllt. Einfache Subjekt- und Objekt-Vorkommen liefern dasselbe Ergebnis wie Werte, die nur transitive Verben aufnehmen. In Tabelle 8 (S. 30) die entsprechenden Ergebnisse der Klassifikation mit transitiven Verben.

		BELEBT			UNBELEBT		
Accuracy	**MCC**	Precision	Recall	F-Score	Precision	Recall	F-Score
0.8	0.16	0.1	0.5	0.17	0.97	0.81	0.89

Tabelle 8: Ergebnisse Klassifizierung Experiment 2 mit transitiven Verben

Es kann jedoch angefügt werden, dass sich der F-Score für BELEBT etwas verbessert, da der Recall sich bemerkbar erhöht. Das ist grundsätzlich positiv, da die Klassifizierung als BELEBT offenbar sehr viel schwieriger ist als die Klassifizierung von UNBELEBT. Jedoch ist die Veränderung so insignifikant, dass nicht wirklich von einer Verbesserung gesprochen werden kann.

5.2.4 Experiment 3: Klassifizierung für hochfrequente Nomen

Ein möglicher Grund für die schlechte Performanz des Klassifizieres liegt in der geringen Menge der Daten begründet. Sobald ein Nomen nur etwa 1-mal auftritt, liegen die Werte für SUBJ, OBJ und VF oftmals ausschließlich bei 0 oder 1. Dies ist oft so willkürlich verteilt, je nachdem, ob die Nomen beim Auftreten in Subjekt- oder Objektposition stehen, dass der Klassifizierer aus diesen Werten keine vernünftigen Regeln ableiten kann. Deshalb wurde ein weiteres Experiment mit hochfrequenten Nomen durchgeführt, deren Frequenz bei > 10, > 50, bzw. > 100 liegt. Da die Datenmenge zu gering war, wurde auf ein Testset verzichtet und nur eine 10-Crossvalidierung durchgeführt, um wenigstens so eine Robustheit testen zu können. Die Ergebnisse finden sich in Tabelle 9 (S. 30).

Frequenz			BELEBT			UNBELEBT		
	Accuracy	**MCC**	Precision	Recall	F-Score	Precision	Recall	F-Score
> 10	0.83	0.38	0.32	0.67	0.44	0.96	0.84	0.9
> 50	0.84	0.37	0.27	0.71	0.39	0.97	0.85	0.91
> 100	0.87	0.47	0.28	0.92	0.43	0.99	0.87	0.93

Tabelle 9: Ergebnisse Klassifizierung Experiment 3 mit hochfrequenten Nomen

Es zeigt sich, dass die Ergebnisse durch die Beschränkung auf hochfrequente Nomen die Ergebnisse durchaus deutlich verbessern können. Für Nomen mit einer Frequenz > 100 konnte der wichtige MCC-Wert aus Experiment 2 um 0.31 Punkte auf 0.47 verbessert werden, hat sich also weit von einer Zufallsklassifikation entfernt. Jedoch entscheidet der Klassifizierer in diesem Fall nur noch anhand des Featues SWMASK. In den Daten der Nomen mit Frequenz > 100 sind relativ viele schwache Maskulina vorhanden, sodass es für den Klassifizierer ausreicht, alle schwachen Maskulina als belebt zu kategorisieren, um ein derartiges Ergebnis zu erreichen.

5.3 Fehlerhafte Klassifizierungen und mögliche Ursachen

Eine erste Fehlerquelle liegt in der Erzeugung der Golddaten. Die Annahme, dass alle Lemmata in den GermaNet-Synsets *Mensch* und *Tier* zwangsläufig belebte Entitäten enthalten, ist stark vereinfacht. Beispielsweise enthält das Synset *Mensch* auch Lemma-Einträge wie *Toter* und das Synset *Tier* Einträge wie *Amöbe* oder *Bandwurm*. Diese Tiere sollten auf einer sprachlichen Belebtheitsskala keinen hohen Rang haben und könnten zu entsprechenden Irritationen in der automatischen Klassifikation führen. Zum einen sollte also überlegt werden, ob Tiere als unbelebte Nomen klassifiziert werden sollten und menschlichen Entitäten als prototypischen belebten Nomen so mehr Gewicht zukommen sollte. Zum anderen bietet es sich an, die Hyponym/Hyperonym-Beziehung von GermaNet einzubeziehen und so innerhalb von GermaNet schon eine Klassifizierung an belebten und unbelebten Nomen vorzunehmen. Ein entsprechender Ansatz wird in Orăsan und Evans (2001) vorgestellt, allerdings benötigt der Ansatz nach Belebtheit annotierte Terminale der Hyperonym-Hierarchien. Auch die Methode, nach Übereinstimmungen im rechten Kontext der Wörter zu schauen, ist nicht immer akkurat, da so auch z.B. *Nikolaisaal* als BELEBT klassifiziert wird (da eine Übereinstimmung mit *Aal* vorliegt). Dies zu umgehen würde eine Erkennung von Komposita-Grenzen erfordern, was ein Gebiet für sich ist und eigene größere Schwierigkeiten bereithält.

Desweiteren existieren in Texten zahlreiche Fälle an belebten Nomen, die nicht durch Lemma-Eintrage aus GermaNet abgefangen werden können. Eine große Klasse stellen hier vor allem aus Adjektiven derivierte Nomen dar wie *Alter*, *Starker* oder auch *24-jähriger*. Um diese Fälle in den Golddaten als BELEBT zu erkennen, ist wohl eine manuelle Belebtheits-Annotation nötig,

die kontextuelle Urteile fällt. Eine auf den Kontext bezogene Datenlage würde natürlich auch neue Möglichkeiten für die Klassifikation an sich bieten.

Es liegt der Verdacht nahe, dass die Datengrundlage TüBa-D/Z nicht genug Datensätze in Form von Sätzen und Token enthält, um anhand von Frequenzen eine korrekte Entscheidung vor allem der Feature SUBJ und OBJ vorzunehmen. Sämtliche vorgestellte Literatur zu dem Thema, wenn auch für andere Sprachen, zeigt eine deutliche Tendenz dafür auf, dass das Einbeziehen von Subjekt- und Objekt-Merkmalen gute Ergebnisse für eine Belebtheitsklassifikation liefert. Es sollte deshalb versucht werden, aus größeren Datenmengen erneut entsprechende Frequenzen zu extrahieren und die Experimente erneut mit den neuen Daten durchzuführen.

5.4 Einordnung der Ergebnisse in den Forschungsdiskurs

Da keine Arbeit zur automatischen Belebtheitsklassifikation im Deutschen vorliegt, ist ein Vergleich mit den Arbeiten zu anderen Sprachen natürlich schwierig. Da man jedoch davon ausgehen kann, dass das Problem für alle Sprachen, in denen Belebtheit nicht offensichtlich über grammatische Merkmale ablesbar ist, sehr ähnlich sein dürfte, soll trotzdem eine entsprechende Einordnung vorgenommen werden.

Die hier sehr kritisch aufgenommenen Ergebnisse der Klassifikation speisen sich aus einem hohen Anspruch an die Robustheit und die Analyse des Klassifizierers, den andere Arbeiten so oft nicht vornehmen. Gerade Øvrelid (2006), deren Klassifikations-Problem grundsätzlich dem vorliegenden ähnelte und deren Ansatz deswegen stark gefolgt wurde, nimmt eine sehr optimistische Baseline von 50% an. Es kann jedoch davon ausgegangen werden, dass die Verteilung belebter und unbelebter Nomen im von Øvrelid (2006) verwendeten Korpus ähnlich aussehen sollte. Die Ergebnisse der vorliegenden Arbeit und die Ergebnisse aus Øvrelid (2006) ähneln sich also sehr stark, auch wenn Øvrelid (2006) die Ergebnisse entsprechend positiver darstellt. In Øvrelid (2009) wird ein Verfahren mit ähnlichen bis identischen Feature vorgestellt, allerdings werden hier statt Entscheidungsbäumen *k-nearest-neighbor*-Ansätze verwendet. Eventuell könnte also ein anderer Klassifizierer die vorliegenden Ergebnisse verbessern.

Orăsan und Evans (2001) und Orăsan und Evans (2007) erreichen sehr hohe F-Score-Werte sowohl für belebte als auch unbelebte Nomen. Ihr lemmabasierter Ansatz, über WordNet und aus dessen Hyponomierelationen Belebtheit

zu lernen, kann also als sehr erfolgreich angesehen werden. Allerdings erfordert dieser Ansatz eine hohe Startmenge an manuell annotierten Daten, um darauf aufbauend hyponyme Belebtheit zu lernen. Ihr Ansatz ist also nicht direkt mit dem vorliegenden Ansatz vergleichbar, da hier die lexikalische Ressource GermaNet nur für die Golddaten-Extraktion genutzt wurde. Für zukünftige Untersuchungen auf Lemma-Basis sollte jedoch darüber nachgedacht werden, ihre Methode für das Deutsche zu adaptieren und zu verfeinern.

Auch Bloem und Bouma (2013) machen eine ähnliche Erfahrung mit Subjekt-Objekt-Verhältnissen wie der vorliegende Ansatz, da sie kaum über die Baseline von 80% (jedes Nomen als UNBELEBT klassifiziert) hinauskommen. Die restlichen Ansätze sind kontext- und tokenbezogen und lassen sich deshalb schlecht mit der vorliegenden Arbeit vergleichen. Jedoch sollte in jedem Fall in Erwägung gezogen werden, in weiterer Forschung auch für das Deutsche kontextbezogene Klassifizierungen vorzunehmen. Besonders Karsdorp et al. (2015) erreichen hier sehr überzeugende Ergebnisse mit relativ geringem Aufwand, eine entsprechende manuelle Annotation von kontextueller Belebtheit vorausgesetzt.

Insgesamt ist durch Hinzunahme dieser Arbeit in den bestehenden Forschungsdiskurs deutlich geworden, dass ein lemmabasierter Ansatz für Belebtheitsklassifikation problematischer ist als man zunächst denke könnte. Es müssen nicht nur Subjekt- und Objekt-Frequenzen einbezogen werden, offensichtlich ist es entscheidend, auch Verb-Semantiken mit einzubeziehen.

5.5 Zukünftige Untersuchungen

Da die Klassifizierung insgesamt nicht sehr erfolgreich war, wenn eine Baseline von 80% angenommen wird und hochfrequente Nomen ausgeklammert werden, ist eine Aufstellung der zukünftigen Forschungsaufgaben umso wichtiger. Es wurden bereits einige Punkte genannt, die hier jedoch noch einmal zusammengefasst und ergänzt werden sollen.

- Ermitteln von Verb-Semantik, die auf Belebtheit hindeutet
- Evaluation für verschiedene Klassifikationsarten
- Verbesserung der Golddaten-Extraktion
- Entwicklung von Ansätzen zu einer kontextbezogenen Klassifikation

Insbesondere die kontextbezogene Klassifikation sollte in naher Zukunft auf den vorgestellten Daten vereinfacht werden, da momentan das TüBa-D/Z-Korpus mit den Bedeutungsklassen aus GermaNet direkt annotiert wird. Zum jetzigen Zeitpunkt sind nicht genug Sätze annotiert, um sie sinnvoll verwenden zu können. Ist die Überführung aus GermaNet jedoch abgeschlossen, sollte eine kontextbasierte Klassifikation jedoch möglich sein.

Besonders vielversprechend scheint auch das Ermitteln von Verbsemantik zu sein. Das Wissen, welche Art von Argumenten in Bezug auf Belebtheit ein transitives Verb fordert, zusammengenommen mit der Frequenz für transitive Subjekte und Objekte sollte die Klassifizierungsgenauigkeit wesentlich steigern und recht schnell von einer Baseline-Klassifikation wegführen. Eventuell kann der sehr elaborierte und komplexe Ansatz aus Baker und Brew (2010) entsprechend adaptiert werden.

6 Zusammenfassung

Es zeigte sich in den vorgestellten Untersuchungen, dass die ursprünglich angenommen Feature nicht die Performanz für eine Belebtheitsklassifikation liefern wie erwartet. Die vorgestellte Methode lag etwa auf Baseline-Niveau, abgesehen von einer Klassifikation hochfrequenter Nomen. Offenbar ist es essenziell notwendig, nicht nur die relativen Subjekt-Objekt-Frequenzen einzubeziehen, sondern diese Frequenzen auch in Zusammenhang mit den regierenden Verben und deren Semantik zu stellen. Nur das Feature SWMASK lieferte positive Ergebnisse, vor allem für hochfrequente Nomen. Es sollte also als deutsch-spezifisches Merkmal in jedem Fall in weitere Untersuchungen aufgenommen werden. Die übrigen Feature SUBJ, OBJ und VF erwiesen sich nicht als völlig unbrauchbar, sondern zeigten im erzeugten Entscheidungsbaum durchaus Tendenzen, die erwarteten Eigenschaften zu besitzen. Sie sollten also nicht völlig verworfen, sondern modifiziert und verbessert werden.

Es hat sich ebenso gezeigt, dass die Baseline für eine Belebtheitsklassifikation im Deutschen recht hoch ist und eine Extraktion von relativ simplen Feature nicht ausreicht, eine befriedigende Accuracy zu erhalten. Dies ist dem hohen Anteil an unbelebten Nomen geschuldet, die eine Klassifikation schnell zu falschen Schlüssen führt. Der lemmabasierte Ansatz ist deshalb weiter auszubauen.

Da für eine automatische Klassifizierung mittels maschinellem Lernen die Auswahl der richtigen Feature entscheidend ist, sollten hier verstärkt weitere Untersuchungen vorgenommen werden, z.B. über die angesprochenen Verb-Semantiken. Außerdem erscheint es sinnvoll, in Zukunft auch kontextbezogene Klassifikationen vorzunehmen, da Belebtheit immer kontextabhängig ist und ein lemmabasierter Ansatz der Problematik deswegen nie ganz gerecht werden kann. Ein kontextbezogener Ansatz erfordert jedoch ein höheres Maß an Ressourcen und manuell annotierten Daten.

Es wurde dennoch ein erster Schritt in Richtung automatische Belebtheitsklassifikation im Deutschen getan, die bei Erfolg sich für zahllose NLP-Anwendungen als nützlich und bereichernd erweisen kann und wird.

Literatur

Baker, K. & Brew, C. (2010). Multilingual animacy classification by sparse logistic regression. *Information Concerning OSDL Ohio State Dissertations in Linguistics*, 52–74.

Berman, J. (2000). *Topics in the clausal syntax of German* (Dissertation). Universität Stuttgart.

Bird, S., Loper, E. & Klein, E. (2009). *Natural language processing with Python*. O'Reilly Media Inc.

Bloem, J. & Bouma, G. (2013). Automatic animacy classification for Dutch. *Computational Linguistics in the Netherlands Journal*, *3*, 82–102.

Bowman, S. R. & Chopra, H. (2012). Automatic animacy classification. In *NAACL-HLT 2012* (S. 7–10).

Eisenberg, P. (2013). *Grundriss der deutschen Grammatik* (4. Aufl., Bd. 1: Das Wort). Stuttgart: Metzler.

Evans, R. & Orăsan, C. (2000). Improving anaphora resolution by identifying animate entities in texts. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000)* (S. 154–162). Lancaster, UK.

Fellbaum, C. (Hrsg.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Foley, W. A. & Van Valin, R. D., Jr. (1985). Information packaging in the clause. In T. Shopen (Hrsg.), *Language typology and syntactic description* (Bd. 1: Clause Structure). Cambridge: Cambridge University Press.

Hamp, B. & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.* Madrid.

Henrich, V. & Hinrichs, E. (2010). GernEdiT - The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Recources and Evaluation (LREC 2010)* (S. 2228–2235). Valletta, Malta.

Karsdorp, F. B., van der Meulen, M., Meder, T. & van den Bosch, A. (2015). Animacy detection in stories. In M. Finlayson, B. Miller, A. Lieto & R. Ronfard (Hrsg.), *Proceedings of the Workshop on Computational Models of Narrative (CMN'15)* (S. 82-97). OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany.

Köpcke, K.-M. (2005). „Die Prinzessin küsst den Prinz“ — Fehler oder gelebter Sprachwandel? *Didaktik Deutsch*, *18*, 67–83.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Matthews, B. M. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, *405*, 442–451.

Orăsan, C. & Evans, R. (2001). Learning to identify animate references. In W. Daelemans & R. Zajac (Hrsg.), *Proceedings of CoNLL-2001* (S. 129–136). Toulouse, France.

Orăsan, C. & Evans, R. (2007). NP animacy identification for anaphora resolution. *Journal of Artificial Intelligence Research*, *29*, 79–103.

Øvrelid, L. (2006). Towards robust animacy classification using morphosyntactic distributional features. In *Proceedings of EACL 2006 Student Research Workshop* (S. 47–54). Trento, Italy.

Øvrelid, L. (2009). Empirical evaluation of animacy annotation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL).*

Øvrelid, L. & Nivre, J. (2007). When word order and part-of-speech tags are not enough — Swedish dependency parsing with rich linguistic features. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)* (S. 447–451).

Schiller, A., Teufel, S., Stöckert, C. & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)* (Bericht). Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Silverstein, M. (1976). Hierarchy of features and ergativity. In R. M. W. Dixon (Hrsg.), *Grammatical categories in Australian languages.* Canberra: Australian Institute of Aboriginal Studies. (zitiert in: Bloem, J. & Bouma, G. (2013). Automatic animacy classification for Dutch. *Computational Linguistics in the Netherlands Journal, 3*, 82–102)

Telljohann, H., Hinrichs, E. W., Kübler, S., Zinsmeister, H. & Beck, K. (2012). *Stylebook for the Tübingen treebank of written German (TüBa-D/Z)* (Bericht). Tübingen: Seminar für Sprachwissenschaft, Universität Tübingen.

Yamamoto, M. (1999). *Animacy and reference. A cognitive approach to corpus linguistics*. Amsterdam, Philadelphia: John Benjamins B.V.

Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., ... Wasow, T. (2004). Animacy encoding in English: Why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation* (S. 118–125). Stroudsburg, PA, USA: Association for Computational Linguistics.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang: Konfusionsmatrizen und generierte Regeln der C5.0-Klassifizierung

6.1 Experiment 1: Binäre Feature

Training: 34134 cases

(a)	(b)	<- classified as
336	6456	(a) animate
128	27205	(b) inanimate

Test: 8532 cases

(a)	(b)	<- classified as
77	1635	(a) animate
18	6802	(b) inanimate

Rule set:

```
Rule 1: (464/128, lift 3.6)
        swmask = t
        ->  class animate  [0.723]

Rule 2: (27729/3627, lift 1.1)
        swmask = f
        ->  class inanimate  [0.869]
```

6.2 Experiment 2: Gewichtung des SUBJ- und OBJ-Features

Experiment 2.1: Alle Vorkommen von SUBJ und OBJ mit relativen Frequenzen

Training: 34134 cases

(a)	(b)	<- classified as
446	6312	(a) animate
233	27143	(b) inanimate

Test: 8532 cases

(a)	(b)	<- classified as
103	1652	(a) animate
56	6721	(b) inanimate

Attribut Usage:
99% OBJ
83% SWMASK
81% SUBJ

Decision tree:

```
swmask = t: animate (541.2/163.3)
swmask = f:
```

```
:...obj > 0.1123596: inanimate (6185.5/680.6)
    obj <= 0.1123596:
    :...subj <= 0.1690141: inanimate (20813.1/3921.3)
        subj > 0.1690141:
        :...subj <= 0.4827586:
            :...subj <= 0.3703704: inanimate (1339.8/386.4)
            :   subj > 0.3703704: animate (208.5/109.9)
            subj > 0.4827586:
            :...obj <= 0.02173913: inanimate (5022.9/1279.3)
                obj > 0.02173913: animate (22.9/9)
```

Rule set:

```
Rule 1: (27746/3591, lift 1.1)
        swmask = f
        ->  class inanimate  [0.871]

Rule 2: (447/114, lift 3.8)
        swmask = t
        ->  class animate  [0.744]

Rule 3: (224/110, lift 2.6)
        subj > 0.3703704
        subj <= 0.4827586
        obj <= 0.1123596
        ->  class animate  [0.509]

Rule 4: (24/9, lift 3.1)
        subj > 0.4827586
        obj > 0.02173913
        obj <= 0.1123596
        ->  class animate  [0.615]

Default class: inanimate
```

Experiment 2.1 - 10-fache Crossvalidierung

Training: 34134 cases
Test: 8532 cases

(a)	(b)	<- classified as
567	7946	(a) animate
325	33828	(b) inanimate

Experiment 2.2 - Nur transitive SUBJ und OBJ

Training: 34134 cases

(a)	(b)	<- classified as
745	6056	(a) animate
759	26574	(b) inanimate

Test: 8532cases

(a)	(b)	<- classified as
175	1537	(a) animate
173	6647	(b) inanimate

Attribut Usage:
99% SUBJ
83% SWMASK
13% OBJ
3% VF

Decision tree:

```
swmask = t: animate (561.8/179.1)
swmask = f:
:...subj <= 0.1716418: inanimate (29324.6/5116.3)
    subj > 0.1716418:
    :...obj > 0.1791045: inanimate (412.4/58.8)
        obj <= 0.1791045:
        :...obj <= 0.005524862:
            :...subj > 0.4736842: inanimate (2686.2/818.4)
            :   subj <= 0.4736842:
            :   :...subj > 0.3581395: animate (74.7/28)
            :       subj <= 0.3581395:
            :       :...vf <= 0.3529412: animate (737.9/489.4)
            :           vf > 0.3529412: inanimate (70.8/14.9)
            obj > 0.005524862:
            :...vf > 0.3157895:
                :...subj <= 0.2857143: inanimate (12)
                :   subj > 0.2857143:
                :   :...obj <= 0.06896552: animate (2)
                :       obj > 0.06896552: inanimate (12/2)
```

```
                vf <= 0.3157895:
                :...subj > 0.375: animate (15.9)
                    subj <= 0.375:
                    :...obj <= 0.0775862: animate (86.9/37)
                        obj > 0.0775862:
                        :...vf <= 0.2380952: animate (119.9/76)
                            vf > 0.2380952: inanimate (17/1)
```

Rule set:

```
Rule 1: (464/128, lift 3.6)
        swmask = t
        ->  class animate  [0.723]

Rule 2: (98/33, lift 3.3)
        subj > 0.3581395
        subj <= 0.4736842
        obj <= 0.1791045
        ->  class animate  [0.660]

Rule 3: (299/154, lift 2.4)
        subj > 0.1716418
        obj > 0.005524862
        obj <= 0.1791045
        ->  class animate  [0.485]

Rule 4: (1219/728, lift 2.0)
        subj > 0.1716418
        subj <= 0.4736842
        obj <= 0.1791045
        ->  class animate  [0.403]

Rule 5: (5934/573, lift 1.1)
        subj <= 0.375
        obj > 0.0775862
        swmask = f
        ->  class inanimate  [0.903]

Rule 6: (454/46, lift 1.1)
```

```
        subj <= 0.2857143
        obj > 0.005524862
        vf > 0.3157895
        ->  class inanimate  [0.897]

Rule 7: (27729/3627, lift 1.1)
        swmask = f
        ->  class inanimate  [0.869]

Default class: inanimate
```

Experiment 2.2 - 10-fache Crossvalidierung

Training: 34134 cases
Test: 8532 cases

(a)	(b)	<- classified as
725	7788	(a) animate
403	33750	(b) inanimate

6.3 Experiment 3: Klassifizierung für hochfrequente Nomen

Frequenz > 10 - 10-fache Crossvalidierung

Training: 2859 cases

(a)	(b)	<- classified as
192	401	(a) animate
93	2173	(b) inanimate

Rule set:

```
Rule 1: (119/12, lift 4.3)
        swmask = t
        ->  class animate  [0.893]

Rule 2: (2654/436, lift 1.1)
        swmask = f
        ->  class inanimate  [0.835]
```

```
Default class: inanimate
```

Frequenz > 50 - 10-fache Crossvalidierung

Training: 576 cases

(a)	(b)	<- classified as
30	80	(a) animate
12	454	(b) inanimate

Rule set:

```
Rule 1: (31/5, lift 4.3)
        swmask = t
        ->  class animate  [0.818]

Rule 2: (543/82, lift 1.0)
        swmask = f
        ->  class inanimate  [0.848]

Default class: inanimate
```

Frequenz > 100

Training: 235 cases

(a)	(b)	<- classified as
11	28	(a) animate
1	195	(b) inanimate

Rule set:

```
Rule 1: (12/1, lift 5.2)
        swmask = t
        ->  class animate  [0.857]

Rule 2: (223/28, lift 1.0)
        swmask = f
        ->  class inanimate  [0.871]

Default class: inanimate
```

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die Arbeit selbständig angefertigt, außer den im Quellen- und Literaturverzeichnis sowie in den Anmerkungen genannten Hilfsmitteln keine weiteren benutzt und alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, unter Angabe der Quellen als Entlehnung kenntlich gemacht habe.

Ort, Datum Unterschrift