

Korpusbasierte Lexikonerstellung mittels nicht paralleler Texte

Reinhard RAPP

Die Extraktion von Wortübersetzungen aus Korpora wurde bislang meist als Zuordnungsproblem betrachtet: Ausgehend von parallelen Korpora wurde ermittelt, welche Sätze bzw. Wörter des Ausgangstextes mit welchen Sätzen bzw. Wörtern der zugehörigen Übersetzung korrespondieren. Obwohl solche Zuordnungsalgorithmen recht gut funktionieren, besteht in der Praxis das Problem, daß für ein gewünschtes Sprachpaar und eine gewünschte Textsorte nur selten ein ausreichend großes paralleles Korpus zur Verfügung steht. Aus diesem Grunde ist es wünschenswert, die Übersetzungen von Wörtern auch ausgehend von nicht parallelen Texten bestimmen zu können. Diese Aufgabe ist sehr viel schwieriger, weil die für parallele Texte wirkungsvollsten sprachstatistischen Indikatoren, nämlich die Wortreihenfolge und die Worthäufigkeit, im Falle nicht paralleler Texte ungeeignet sind. In der vorliegenden Arbeit wird deshalb zunächst ein neuer statistischer Indikator eingeführt, der auf der Annahme beruht, daß das gemeinsame Auftreten der Wörter in Texten verschiedener Sprachen eine Korrelation aufweist. Anschließend wird ein Algorithmus vorgestellt, der mit Hilfe dieses Indikators die Übersetzungen von Wörtern auf der Basis nicht paralleler Korpora mit einer Trefferquote von etwa 70% bestimmen kann.

Wortkookkurrenzen in verschiedensprachigen Texten

Das Problem der Identifikation von Wortübersetzungen, das hier mit maschinellen Methoden angegangen werden soll, stellt sich Übersetzern und Dolmetschern bei ihrer täglichen Arbeit. Zur terminologischen Vorbereitung von Übersetzungsaufträgen in einem neuen Gebiet ist es für diese selbstverständlich, thematisch relevante Texte in den benötigten Sprachen durchzuarbeiten. Obwohl es sich bei diesen Texten in der Regel nicht um Ausgangstexte und zugehörige Übersetzungen handelt, ist es den Über-

setzern dennoch in vielen Fällen möglich, aus dem vorgefundenen Sprachgebrauch Rückschlüsse auf die korrekten Übersetzungen von Fachtermini zu ziehen. In der vorliegenden Arbeit wird eine maschinelle Methode vorgestellt, die genau dieses Problem angeht. Auch wenn hierbei kein wirkliches Sprachverstehen realisiert wird, so beruht der Ansatz doch auf den Grundlagen des klassischen Assoziationismus und weist damit eine sehr viel höhere gedächtnispsychologische Plausibilität auf als die bislang verwendeten Alignment-Verfahren (siehe hierzu WETTLER et al. 1993; RAPP 1996). In der vorliegenden Arbeit liegt der Schwerpunkt jedoch nicht auf kognitiven Aspekten, sondern auf der Beschreibung des tatsächlich implementierten Algorithmus.

Grundlage des Algorithmus ist die Annahme, daß verschiedene Sprachen ähnliche Kookkurrenzmuster aufweisen. Wenn also beispielsweise in einem deutschen Text die beiden Wörter *Lehrer* und *Schule* sehr viel häufiger als rein statistisch zu erwarten gemeinsam auftreten, so ist zu vermuten, daß in einem englischen Text die Übersetzungen dieser beiden Wörter, also *teacher* und *school*, ebenfalls überzufällig häufig zusammen vorkommen. Dies ist für parallele Texte unmittelbar plausibel. In einer Fallstudie haben RAPP et al. (1995) jedoch gezeigt, daß diese Annahme – obgleich in geringerem Maße – auch für nicht parallele Texte gültig ist.

Da die Wortkookkurrenzen in nicht parallelen Korpora aber ein wesentlich schwächerer sprachstatistischer Indikator sind als die Wortreihenfolge in parallelen Korpora, werden größere Korpora benötigt, und die Auswahl geeigneter statistischer Methoden ist von besonderer Wichtigkeit. Andererseits garantiert das Verfahren eine größere Robustheit, da die beim Alignment paralleler Korpora auftretenden Probleme mit Änderungen in der Textabfolge sowie mit Auslassungen und Einfügungen von Textpassagen von vornherein vermieden werden.

Bedeutungsähnlichkeiten von Wörtern

Nach RUGE (1995) läßt sich die Bedeutungsähnlichkeit von Wörtern dadurch berechnen, daß ihre Übereinstimmung in bezug auf die in ihrer Umgebung auftretenden Wörter bestimmt wird. So könnte man etwa die semantische Verwandtschaft der beiden Wörter *rot* und *blau* daraus ablei-

ten, daß beide häufig im Zusammenhang mit Begriffen wie *Farbe, Blume, Lackierung, Kleid, Auto, hell, dunkel, leuchtend, schön* etc. verwendet werden. Wird für jedes in einem Korpus vorkommende Wort ein Kookkurrenzvektor angelegt, in dem die Häufigkeiten des gemeinsamen Auftretens mit allen anderen Wörtern gespeichert sind, so lassen sich die Bedeutungsähnlichkeiten zwischen Wörtern auf einfache Vektorvergleiche zurückführen. Sollen die zu einem bestimmten Wort bedeutungsähnlichsten anderen Wörter ermittelt werden, so werden mittels eines geeigneten Ähnlichkeitsmaßes die Ähnlichkeiten zwischen dem Vektor dieses Wortes und den Vektoren aller anderen Wörter berechnet. Diejenigen Wörter, für die sich die höchsten Ähnlichkeitswerte ergeben, sollten dann auch die höchste Bedeutungsähnlichkeit aufweisen. Praktische Realisierungen dieser Methode haben zu ausgezeichneten Ergebnissen geführt (GREFENSTETTE 1994; RUGE 1995; AGARWAL 1995; LIN 1998).

Der hier vorgestellte Ansatz stellt die Übertragung dieser Vorgehensweise vom einsprachigen auf den zweisprachigen Fall dar. Vorausgesetzt werden ein ausgangssprachliches und ein zielsprachliches Korpus sowie ein von Beginn an zur Verfügung stehendes kleineres Basislexikon. Angestrebt wird die Erweiterung dieses Basislexikons. Die Vorgehensweise ist nun wie folgt: Ausgehend von einem Korpus der Zielsprache wird eine Kookkurrenzmatrix ausgezählt, deren Zeilen dem Vokabular im Korpus und dessen Spalten dem zielsprachlichen Vokabular im Basislexikon entsprechen. Soll nun die Übersetzung eines unbekanntes ausgangssprachlichen Wortes bestimmt werden, so wird zunächst auf der Basis des ausgangssprachlichen Korpus der Kookkurrenzvektor dieses Wortes ausgezählt. Anschließend werden unter Verwendung des Basislexikons die in diesem Vektor auftretenden Wörter übersetzt. Alle im Basislexikon nicht gefundenen Wörter werden gelöscht. Der resultierende Vektor wird mit allen Vektoren in der zielsprachlichen Kookkurrenzmatrix verglichen. Dasjenige Wort, für das sich die höchste Ähnlichkeit ergibt, wird als Übersetzung des unbekanntes Wortes betrachtet.

Diese Vorgehensweise sei anhand eines Beispielles näher erläutert, das auf dem Sprachpaar Deutsch/Englisch basiert. Dabei ist Deutsch die Ausgangssprache und Englisch die Zielsprache. Das Basislexikon umfasse

folgende Einträge (22 000 Stichwörter im tatsächlich verwendeten Lexikon):

<i>Deutsch</i>	<i>Englisch</i>
blau	blue
grün	green
Farbe	color
Junge	boy
Klasse	class
Schule	school

Zunächst wird auf der Grundlage des englischen Korpus die englische Kookkurrenzmatrix berechnet. Es handelt sich dabei um eine asymmetrische Matrix, deren Zeilen alle Wörter enthält, die im englischen Korpus vorkommen, deren Spalten aber nur diejenigen Wörter umfaßt, die in der zielsprachlichen Spalte des Basislexikons auftreten.

	blue	boy	color	class	green	school
black	1	0	1	0	1	0
blue	0	0	1	0	1	0
boy	0	0	0	1	0	1
color	1	0	0	0	1	0
class	0	0	0	0	0	1
green	1	0	1	0	0	0
ocean	1	0	0	0	0	0
pupil	0	1	0	1	0	1
school	0	0	0	1	0	0
teacher	0	0	0	1	0	1
white	1	0	1	0	1	0

Zur Vereinfachung werden in der Kookkurrenzmatrix lediglich zweiwertige Einträge angenommen. Eine 1 bedeutet dabei, daß das zugehörige Wortpaar signifikant häufiger gemeinsam auftritt, als dies statistisch zu erwarten wäre. Eine 0 steht hingegen für zufällig häufiges gemeinsames Auftreten.

Ziel sei es nun, die englische Übersetzung des Wortes *Schüler*, das im Basislexikon nicht enthalten ist, zu bestimmen. Hierzu wird zunächst auf der Grundlage des deutschen Korpus der Kookkurrenzvektor dieses Wortes bestimmt:

	blau	Farbe	Ferien	grün	Junge	Klasse	Pflanze	Schule
Schüler	0	0	1	0	1	1	0	1

Unter Verwendung des Basislexikons wird dieser Vektor in das Englische übersetzt:

	blue	color	???	green	boy	class	???	school
Schüler	0	0	1	0	1	1	0	1

Die Wörter *Ferien* und *Pflanze* konnten dabei nicht übersetzt werden, da sie nicht im Basislexikon enthalten sind. Die zugehörigen Vektorpositionen haben deshalb keinen Informationsgehalt und werden einfach gelöscht. Der resultierende Vektor wird alphabetisch sortiert, so daß die Wortreihenfolge der in der bereits berechneten englischen Kookkurrenzmatrix entspricht:

	blue	boy	color	class	green	school
Schüler	0	1	0	1	0	1

Dieser Vektor wird nun nacheinander mit jedem der Vektoren in der englischen Matrix verglichen. Als Ähnlichkeitsmaß kann beispielsweise die Anzahl der jeweils übereinstimmenden Nullen und Einsen herangezogen

gen werden, d.h. der Ähnlichkeitswert erhöht sich im Falle jeder übereinstimmenden Vektorposition um eins. Man erhält damit den folgenden Ähnlichkeitsvektor:

	Ähnlichkeit
black	0
blue	1
boy	5
color	1
class	4
green	1
ocean	2
pupil	6
school	4
teacher	5
white	0

Da sich für *pupil* der höchste Wert ergibt, würde man also korrekt folgern, daß es sich bei dem Wort *pupil* um die Übersetzung des deutschen Wortes *Schüler* handelt. Auf den nächsten Rängen folgen assoziierte Wörter wie *boy*, *teacher*, *class* und *school*.

Simulation

Der Simulation liegen die Jahrgänge 1993 bis 1996 der *Frankfurter Allgemeinen Zeitung* (135 Millionen Wörter) sowie die Jahrgänge 1990 bis 1994 des *Guardian* (163 Millionen Wörter) zugrunde. Als Basislexikon diente das *Collins Gem German Dictionary* mit etwa 22 000 Stichwörtern. Die beiden Korpora wurden jeweils lemmatisiert und es wurden die Funktionswörter eliminiert. Auf der Basis dieser Korpora wurden eine englische Kookkurrenzmatrix sowie die Kookkurrenzvektoren zu 100 deutschen Testwörtern ausgezählt. Die Auswahl der Testwörter basierte auf einer

Wortliste von RUSSELL (1970). Um signifikante Kookkurrenzen hervorzuheben, wurden sämtliche Kookkurrenzhäufigkeiten mit Hilfe des von DUNNING (1993) vorgeschlagenen *log-likelihood*-Tests in Assoziationsstärken umgerechnet. Der *log-likelihood*-Test ist ähnlich dem Chi-Quadrat-Test ein statistischer Signifikanztest, eignet sich jedoch besser für Ereignisse mit geringer Häufigkeit, die bei korpusbasierten Arbeiten von großer Bedeutung sind (“sparse-data-problem”). Die Berechnung beruht auf folgender Formel:

$$\begin{aligned} -2 \log \lambda &= \sum_{i,j \in \{1,2\}} k_{ij} \log \frac{k_{ij}N}{C_i R_j} \\ &= k_{11} \log \frac{k_{11}N}{C_1 R_1} + k_{12} \log \frac{k_{12}N}{C_1 R_2} + k_{21} \log \frac{k_{21}N}{C_2 R_1} + k_{22} \log \frac{k_{22}N}{C_2 R_2} \end{aligned}$$

hierbei gilt:

$$C_1 = k_{11} + k_{12} \quad C_2 = k_{21} + k_{22}$$

$$R_1 = k_{11} + k_{21} \quad R_2 = k_{12} + k_{22}$$

$$N = k_{11} + k_{12} + k_{21} + k_{22}$$

Mit den Parametern k_{ij} ausgedrückt durch Korpushäufigkeiten:

k_{11} = Häufigkeit des gemeinsamen Auftretens der Wörter A und B

k_{12} = Häufigkeit des Wortes A – k_{11}

k_{21} = Korpushäufigkeit des Wortes B – k_{11}

k_{22} = Anzahl Tokens im Korpus – Häufigk. von A – Häufigk. von B

Anschließend wurde für jeden der 100 deutschen Vektoren mit Hilfe der City-Block-Metrik der jeweils ähnlichste englische Vektor berechnet:

$$s = \sum_{i=1}^n |A_i - B_i|$$

Ergebnisse und Evaluierung

Die Tabelle im Anhang zeigt für alle 100 Testwörter die durch die Simulation bestimmten Übersetzungen, wobei für jedes deutsche Wort die jeweils fünf ähnlichsten englischen Wörter angegeben sind. Zusätzlich wird für jedes Wort die erwartete Übersetzung angegeben. Für insgesamt 65 der 100 Testwörter stimmt das Wort mit der höchsten Ähnlichkeit tatsächlich mit der erwarteten Übersetzung überein. Für weitere sieben Testwörter ist diese Übereinstimmung zwar nicht gegeben, aber das an erster Stelle plazierte Wort stellt dennoch eine korrekte Übersetzung des Ausgangswortes dar. In den übrigen 28% der Fälle wurde vom Programm eine falsche Übersetzung vorhergesagt. Zumeist ist die fälschlicherweise vorhergesagte Übersetzung jedoch nicht völlig abwegig, sondern es handelt sich um ein mit der gesuchten Übersetzung stark assoziiertes Wort. Beispielsweise wird als Übersetzung von *Frau* an erster Stelle *man* angegeben, während *woman* erst an zweiter Stelle folgt. Dieser Fehlertypus ist auf Grund des verwendeten assoziationalistischen Ansatzes zu erwarten.

Ein noch ungelöstes Problem besteht bei mehrdeutigen Wortformen. Diese werden nicht explizit behandelt, sondern es setzt sich im allgemeinen die in den jeweiligen Korpora häufigere Bedeutung durch. Beispielsweise wurden auf das Wort *Kohl* die Übersetzungen *Major*, *Kohl*, *Thatcher*, *Gorbachev* und *Bush* generiert, während die erwartete Übersetzung *cabbage* erst auf einem sehr viel späteren Rangplatz erscheint.

Trotz dieser Schwächen ist es mit der hier vorgestellten Methode gelungen, die in früheren Arbeiten vorgestellten Resultate, die bei ungefähr 30% richtiger Vorhersagen liegen (FUNG & MCKEOWN 1997; FUNG & YEE 1998), signifikant zu verbessern.

Literatur

- AGARWAL, R. (1995): Semantic Feature Extraction from Technical Texts with Limited Human Intervention. Dissertation, Mississippi State University.
- DUNNING, T. (1993): Accurate methods for the statistics of surprise and coincidence. In: *Computational Linguistics* 19/1, 61-74.

- FUNG, P. and MCKEOWN, K. (1997): Finding terminology translations from non-parallel corpora. In: *Proceedings of the 5th Annual Workshop on Very Large Corpora*, Hong Kong: August 1997, 192-202.
- FUNG, P. and YEE, L.Y. (1998): An IR approach for translating new words from nonparallel, comparable texts. In: *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, 1, 414-420. Montreal.
- GREFENSTETTE, G. (1994): *Explorations in Automatic Thesaurus Discovery*. Dordrecht: Kluwer.
- LIN, D. (1998): Automatic Retrieval and Clustering of Similar Words. In: *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, 2, 768-773. Montreal.
- RAPP et al. (1995): R. R., S. ARMSTRONG und M. WETTLER (1995): Die maschinelle Identifikation von Wortübersetzungen in nicht parallelen Texten. In: L. HITZENBERGER (Hrsg.): *Angewandte Computerlinguistik. Vorträge im Rahmen der Jahrestagung der GLDV, Regensburg 1995*, 241-252. Hildesheim: Olms.
- RAPP, R. (1996): *Die Berechnung von Assoziationen*. Hildesheim: Olms.
- RUGE, G. (1995): *Wortbedeutung und Termassoziation*. Hildesheim: Olms.
- RUSSELL, W. A. (1970): The complete German language norms for responses to 100 words from the Kent-Rosanoff word association test. In: L. POSTMAN and G. KEPPEL (eds.): *Norms of Word Association*, 53-94. New York: Academic Press.
- WETTLER et al. (1993): M. W., R. RAPP und R. FERBER, *Freie Assoziationen und Kontiguitäten von Wörtern in Texten*. In: *Zeitschrift für Psychologie* 201, 99-108.

Anhang: Ergebnisse für die 100 Testwörter

Testwort	Erwartete Übersetzung (Rang)	Die fünf stärksten vorhergesagten Übersetzungen
Adler	eagle (924)	player, United, bird, Cardiff, Park
ängstlich	nervous (1)	nervous, know, see, someone, think
arbeiten	work (1)	work, see, know, means, time
Arzt	doctor (1)	doctor, patient, hospital, NHS, specialist
Baby	baby (1)	baby, child, mother, daughter, father
Bad	bath (139)	precious, dinky, seem, nice, known
Bequemlichkeit	convenience (732)	people, someone, means, necessary, life
Berg	mountain (1)	mountain, hill, village, river, desert
Bett	bed (1)	bed, morning, room, sleep, day

Testwort	Erwartete Übersetzung (Rang)	Die fünf stärksten vorhergesagten Übersetzungen
Bibel	Bible (1)	Bible, book, religion, read, text
bitter	bitter (1)	bitter, sad, terrible, concede, humiliate
blau	blue (1)	blue, grey, dark, white, yellow
Blüte	blossom (9)	bloom, flower, plant, phenomenon, bird
Brot	bread (1)	bread, cheese, meat, food, butter
Bürger	citizen (1)	citizen, individual, believe, means, argue
Butter	butter (2)	cheese, butter, carton, bread, sugar
Dieb	thief (1)	thief, steal, burglar, suspect, murderer
dunkel	dark (1)	dark, grey, beautiful, bright, see
durstig	thirsty (13)	lemonade, bottle, litre, slurry, tick
Erde	earth (1)	earth, surface, planet, sea, land
essen	eat (1)	eat, meal, cook, food, meat
Fenster	window (1)	window, lock, open, wall, roof
Fluß	river (1)	river, lake, sea, water, canal
Frau	woman (2)	man, woman, boy, friend, wife
Freude	joy (1)	joy, moment, certainly, occasion, surprise
Fuß	foot (1)	foot, walk, saw, move, run
Gedächtnis	memory (1)	memory, context, worship, history, character
gelb	yellow (1)	yellow, blue, red, pink, green
Gerechtigkeit	justice (10)	security, worker, cohesion, chapter, democrat
Gesundheit	health (3)	concerned, argue, health, believe, means
glatt	smooth (1)	smooth, surface, beneath, scratch, course
grün	green (1)	green, colour, yellow, dark, pink
Hammelfleisch	mutton (5795)	Bushey, Referee, machinery, Forwards, herb
Hammer	hammer (1)	hammer, category, apart, grace, Berlin
Hand	hand (1)	hand, thought, opinion, see, simply
hart	hard (5)	tough, believe, well, strong, hard
Haus	house (1)	house, home, city, see, thought
Häuschen	cottage (2)	bungalow, cottage, house, hut, village
hoch	high (1)	high, low, increase, reduce, lower
hungrig	hungry (1)	hungry, bird, know, maybe, insect
Junge	boy (1)	boy, lady, girl, friend, man
kalt	cold (1)	cold, Gulf, postcold, IranIraq, civil

Testwort	Erwartete Übersetzung (Rang)	Die fünf stärksten vorhergesagten Übersetzungen
Käse	cheese (1)	cheese, sausage, meat, bacon, bread
Kind	child (1)	child, daughter, son, father, mother
Kohl	cabbage (17074)	Major, Kohl, Thatcher, Gorbachev, Bush
kommandieren	command (27)	Knapp, TUC, McAvoy, GMB, Hart
König	King (1)	King, Prince, Queen, George, Charles
Kopf	head (1)	head, thought, hand, heads, simply
Krankheit	sickness (86)	disease, illness, Aids, patient, doctor
kurz	short (43)	first, same, several, long, three
Lampe	lamp (1)	lamp, candle, bulb, light, switch
lang	long (8)	later, ago, after, earlier, same
langsam	slow (1)	slow, quick, begin, time, well
laut	loud (162)	said, believe, yesterday, hear, present
Licht	light (1)	light, moment, idea, surface, seem
Löwe	lion (1)	lion, Arsenal, Liverpool, Villa, side
Mädchen	girl (1)	girl, boy, man, brother, lady
Magen	stomach (2)	lung, stomach, kidney, infection, liver
Mann	man (1)	man, brother, boy, people, lady
Mond	moon (1)	moon, Mundo, Pais, Monde, Welt
Musik	music (1)	music, theatre, musical, dance, song
Nadel	needle (3)	humid, poker, needle, pursuit, tightly
Obst	fruit (1)	fruit, vegetable, meat, rice, pasta
Ofen	oven (2)	heat, oven, stove, house, burn
Ozean	ocean (1)	ocean, subcontinent, Gymkhana, sea, island
pfeifen	whistle (3)	linesman, referee, whistle, blow, offside
Priester	priest (1)	priest, bishop, clergy, church, Pope
Quadrat	square (5)	rectangle, squared, diameter, triangle, square
rauh	rough (496)	icy, prevailing, warmer, chilly, change
Religion	religion (1)	religion, culture, faith, religious, belief
rot	red (3)	yellow, credit, red, identity, greeting
ruhig	quiet (3)	time, see, quiet, calm, thought
Salz	salt (1)	salt, sugar, butter, cook, rice
sauer	sour (922)	acid, monsoon, heavy, millimetre, torrential
Schaf	sheep (1)	sheep, cattle, cow, pig, goat

Testwort	Erwartete Übersetzung (Rang)	Die fünf stärksten vorhergesagten Übersetzungen
Schere	scissors (1)	scissors, keeping, mean, possible, need
schlafen	sleep (1)	sleep, morning, day, bed, thought
Schmetterling	butterfly (6)	freestyle, breast-, backstroke, hurdle, steeplechase
schnell	quick (1)	quick, fast, see, far, simply
schön	beautiful (1)	beautiful, wonderful, thought, see, love
schwarz	black (1)	black, white, dark, grey, wear
schwer	heavy (1)	heavy, serious, seem, see, believe
Soldat	soldier (1)	soldier, army, troop, force, civilian
Sorge	care (418)	fear, worry, concern, concerned, country
Spinne	spider (6)	neural, unguarded, safety, mosquito, forehead
Stadt	city (1)	city, town, village, country, region
Stiel	handle (11667)	lolly, cube, rink, skater, cream
Straße	street (2)	road, street, city, town, walk
Stuhl	chair (1)	chair, desk, table, watch, simply
süß	sweet (1)	sweet, smell, delicious, taste, love
Tabak	tobacco (1)	tobacco, cigarette, consumption, nicotine, drink
Teppich	carpet (1)	carpet, painting, furniture, collection, mirror
tief	deep (1)	deep, seem, see, thought, fact
Tisch	table (1)	table, room, simply, someone, usually
träumen	dream (1)	dream, thought, see, realize, indeed
weich	soft (54)	especially, indeed, add, seem, perhaps
weiß	white (46)	know, thought, see, think, fact
Whisky	whisky (1)	whisky, beer, Scotch, bottle, wine
wünschen	wish (6)	believe, insist, present, means, argue
Zorn	anger (1)	anger, frustration, fear, despair, fact