Heinrich C. Mayr, Martin Pinzger (Hrsg.): INFORMATIK 2016,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2016 2005

Statistisches Lernen wortbasierter Morphologie

Maciej Sumalvico[1]

Abstract: Im Rahmen der Promotion wird ein Ansatz zum automatischen Lernen von Sprachmorphologie entwickelt. Dabei wird auf die Zerlegung von Wörtern in kleinere strukturelle Elemente (die s.g. *Morpheme*) verzichtet und stattdessen mit Transformationsregeln gearbeitet, die an ganzen Wörtern operieren. Das Lernen wird mithilfe eines probabilistischen Modelles realisiert, das mit dem EM-Algorithmus trainiert wird. Die gelernten morphologischen Regeln können bei verschiedenen praktischen Problemen der Sprachtechnologie angewandt werden, um den Umgang mit unbekannten Wörtern zu verbessern.

Keywords: Natural Language Processing, NLP, Sprachtechnologie, Grammatik, Morphologie, maschinelles Lernen, statistische Modellierung.

1 Einleitung

Die Morphologie ist ein Bereich der Sprachgrammatik, der strukturelle Zusammenhänge zwischen Wörtern beschreibt. Dabei wird die *morphologische Analyse* eines Wortes üblicherweise als die Zerlegung in minimale bedeutungs- oder funktionstragende Einheiten, die s.g. *Morpheme*, verstanden [AF11]. Zum Beispiel besteht das im Titel dieses Artikels enthaltene Wort *wortbasierter* aus folgenden Morphemen: *wort-bas-ier-t-er*. Darunter sind neben den Morphemen, die für die Kernbedeutung verantwortlich sind (*wort*, *bas*), auch zwei derivationelle Affixe (*-ier*, *-t*) zu finden, sowie die Flexionsendung *-er*.

In der automatischen Sprachverarbeitung ist ein morphologischer Analysierer ein häufiger Bestandteil von Prozessketten. Da es unmöglich ist, ein vollständiges Lexikon von Wörtern einer Sprache zu erstellen, müssen die Systeme der Sprachtechnologie immer auch mit unbekannten Wörtern umgehen können. Durch morphologische Analyse können die Merkmale eines unbekannten Wortes durch seine strukturelle Beziehung zu bekannten Wörtern erraten werden.

Üblicherweise werden morphologische Analysierer als handgeschriebene Grammatiken und Lexika erstellt und anschließend zu endlichen Automaten kompiliert, die eine performante Durchsuchung ermöglichen. Dieser Ansatz, bekannt als *Zwei-Ebenen-Morphologie*, geht auf die Arbeit von Koskenniemi [Ko83] zurück. Formalismen wie XFST [BK03] ermöglichen die Kompilierung von komplexen Grammatiken und die Optimierung von resultierenden Automaten. Dieser Ansatz zeichnet sich zwar durch hohe Effizienz und Genauigkeit aus, erfordert aber erheblichen Aufwand bei der manuellen Erstellung der Grammatiken und Lexika. Deshalb sind die Methoden des automatischen Lernens von Morphologie ein aktuelles Forschungsthema.

[1] Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, 04109 Leipzig, sumalvico@informatik.uni-leipzig.de

Obwohl fast alle vorhandenen Ansätze zum Lernen von Morphologie auf die Zerlegung von Wörtern in Morpheme zielen [HB11], ist das eigentliche Ziel die Vorhersage von Merkmalen unbekannter Wörter. Die Zerlegung ist dafür weder ausreichend noch notwendig. Darüber hinaus wurde der Begriff des Morphems auch in der Linguistik kritisiert, was zur Formulierung von „morphemlosen" Morphologietheorien geführt hat [An92, FSM97]. Manche Kritikpunkte sind auch für das automatische Lernen von Morphologie relevant: z.B. hat die morphembasierte Analyse Schwierigkeiten mit nichtkonkatenativen Operationen (*Haus*:*Häuser*), sie führt Entitäten ein, die in der Oberflächenform des Wortes nicht zu erkennen sind („Null-Affixe"), die Morphemgrenzen werden durch phonologische Prozesse verwischt, der Sinn der Zerlegung von manchen abgeleiteten Wörtern ist durch verwischte Beziehung zum Grundwort (z.B. *gehören* < *ge-hören*) oder sogar das Fehlen vom Grundwort (*vergessen* < *ver-*gessen*) fraglich. Aus diesen Gründen stellt sich die hier vorgestellte Promotion das Ziel, eine Methode für das automatische Lernen von Morphologie zu entwickeln, die direkt auf die Ableitung neuer Wörter und ihrer Eigenschaften zielt, ohne die Wörter in Morpheme zu zerlegen. Dabei wird insbesondere die *Whole Word Morphology* von [FSM97] als linguistische Grundlage verwendet.

2 Stand der Forschung und aktuelle Trends

Die Aufgabe des automatischen Lernens von Morphologie ist in der Literatur seit langem bekannt: Der erste Ansatz wurde bereits in den 50er Jahren vorgestellt [Ha55]. Lange wurde die Aufgabe vor allem durch heuristische Ansätze gelöst, wie die von Harris eingeführte *Letter Successor Variety* (LSV) [Go06, Bo08] oder die Entdeckung von morphologisch verwandten Wörtern mittels verschiedener Ähnlichkeitsmaße [YW00, BMT02, Ki13]. Ein heuristischer Ansatz, der auf Ganzwortmorphologie (*Whole Word Morphology*) basiert, wurde von Neuvel und Fulop [NF02] vorgeschlagen. Für das überwachte Lernen wurden hingegen die üblichen Methoden des maschinellen Lernens verwendet, wie u.a. Conditional Random Fields [Ru13], Pair Hidden Markov Models [Cl02] oder Maximum Entropy Classifiers [CDvG08].

In den letzten Jahren gewinnen probabilistische Modelle auch beim unüberwachten Lernen immer mehr an Bedeutung [CL05, PCT09, Ca11]. Ihr großer Vorteil ist, dass das gleiche Modell oft sowohl überwacht, als auch unüberwacht trainiert werden kann. Eine zweite wichtige Entwicklung ist die Einbeziehung vom Wortkontext in das Lernen von Morphologie. In früheren Arbeiten wurde dies durch kookkurenzbasierte Ähnlichkeitsmaße erreicht [BMT02, Bo08]. In der letzten Zeit ist mit der Veröffentlichung von `word2vec` [Mi13, MYZ13] ein mächtiges Werkzeug für die numerische Beschreibung von Wortbedeutung verfügbar geworden. Seine Nützlichkeit bei der Aufgabe des Morphologielernens wurde bereits nachgewiesen [SO15].

3 Ein probabilistisches Modell für die Ganzwortmorphologie

Im Rahmen der hier vorgestellten Promotion wird ein generatives probabilistisches Modell vom Sprachlexikon entwickelt, in dem morphologische Zusammenhänge zwischen Wör-

tern mittels Transformationsregeln, die an ganzen Wörtern operieren, ausgedrückt werden. Zum Beispiel wird die Regel, die aus *Haus Häuser* ableitet, folgendermaßen dargestellt:[2]

$$/X_1\mathrm{a}X_2/_{\mathrm{N.SG}} \rightarrow /X_1\ddot{\mathrm{a}}X_2\mathrm{er}/_{\mathrm{N.PL}} \tag{1}$$

Die Objekte innerhalb von $/\cdot/$ sind ganze, existierende Wörter. X_1 und X_2 sind variable Elementen, die mit einer beliebigen Kette von Phonemen (bzw. Buchstaben) instantiiert werden können und von der Regel unverändert bleiben. Zusätzlich zur phonologischen oder ortographischen Repräsentation von Wörtern kann die Regel auch an verschiedenen anderen Merkmalen operieren, wie syntaktische Merkmale (Wortart, Flexionsmerkmale) oder Bedeutung.

Das probabilistische Modell nach aktuellem Stand wurde in [Ja15] detalliert vorgestellt. Es betrachtet Lexika als Graphen (genauer: gerichtete Wälder) von Wörtern mit Ableitungsrelation und schreibt ihnen Wahrscheinlichkeiten zu (Abb. 1). Die Wahrscheinlichkeit eines Wortes, das keine eingehende Kante besitzt, wird aus einer unigrammbasierten Verteilung über alle Buchstabenketten ausgerechnet (Produkt von Häufigkeiten von Buchstaben). Sie ist typischerweise sehr klein und nimmt mit der Länge des Wortes stark ab. Die Wahrscheinlichkeit von Wörtern, die durch Regeln abgeleitet werden (d.h. eine eingehende Kante besitzen), wird für jede Regel festgelegt und als Modellparameter behandelt. So erhalten diejenigen Lexika eine hohe Wahrscheinlichkeit, die oft wiederkehrende strukturelle Zusammenhänge nutzen, um Wörter abzuleiten und die Anzahl von Wurzelknoten zu verringern.

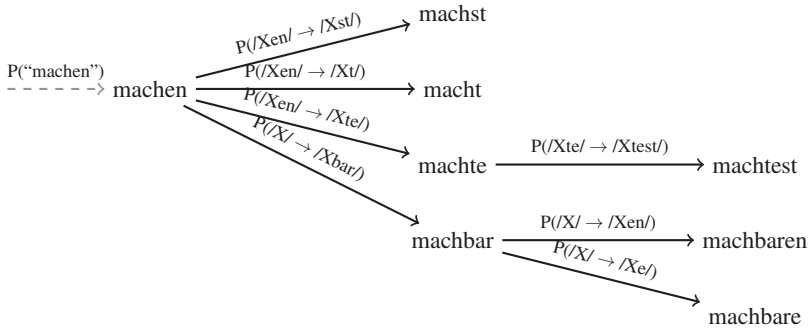

Abb. 1: Fragment eines möglichen deutschen Lexikons.

Das Modell wird im Paradigma der bayesschen Statistik erstellt: Die beobachteten Daten sind das Vokabular V und das zu optimierende Parameter die Regelmenge R. Bei der Berechnung von Wahrscheinlichkeiten spielt auch die unbekannte Menge von Kanten E eine Rolle. Der zu optimierende Wert ist die a-posteriori Wahrscheinlichkeit der Regelmenge

[2] In der ursprünglichen Theorie von [FSM97] sind die Regeln immer bidirektional: sie drücken keine Ableitung aus, sondern eine ungerichtete Relation. In dieser Arbeit wird mit Hinblick auf das probabilistische Modell eine Ableitungsrichtung eingeführt.

gegeben Daten: $P(R|V)$, die folgendermaßen transformiert wird:

$$\begin{aligned} \arg\max_R P(R|V) &= \arg\max_R \frac{P(V|R)P(R)}{P(V)} = \arg\max_R P(V|R)P(R) \\ &= \arg\max_R \sum_E P(V,E|R)P(R) \end{aligned} \tag{2}$$

Neben der Wahrscheinlichkeit von Graphen $P(V,E|R)$, deren Berechnung auf der Abb. 1 schematisch gezeigt wird, wird also auch die Komplexität der gelernten Grammatik durch die a-priori Wahrscheinlichkeit $P(R)$ kontrolliert. Für die Suche nach der optimalen Menge von Regeln können zwei verschiedene Varianten des Expectation-Maximization-Algorithmus [DLR77, SCR12] angewandt werden: entweder wird abwechselnd der Graph und die Regelmenge optimiert („hard EM“), oder aber wird nach einer Regelmenge gesucht, die den Erwartungswert des Log-Likelihoods über alle möglichen Kantenstrukturen maximiert („soft EM“). Die letztere Methode ist im Moment noch in der Entwicklungsphase. Es wird geplant, für die Berechnung von Erwartungswerten Markov Chain Monte Carlo (MCMC) Methoden zu verwenden [RC05]. Das überwachte Trainieren vom Modell ist ebenfalls möglich – in dem Fall wird die Suche nach der optimalen Regelmenge vereinfacht, da die Kantenmenge E bekannt ist.

Der Formalismus lässt sich leicht erweitern, indem Wörter nicht nur als Zeichenketten betrachtet werden, sondern als Vektoren von Merkmalen, die außer ortographischer Repräsentation auch andere Informationen enthalten. Dafür müssen lediglich Wahrscheinlichkeitsverteilungen für die Merkmalswerte der Wurzelwörter und die von Regeln durchgeführten Transformationen festgelegt werden. Zum Beispiel hat sich Worthäufigkeit als ein nützliches Merkmal erwiesen. Das am vielversprechendste Merkmal sind die von `word2vec` berechneten Vektoren, die den Kontext des Wortes – und damit seine Bedeutung – als Koordinaten in einem kontinuierlichen Raum erfassen. Ihre Integration in das Modell wird für die nahe Zukunft geplant.

4 Arbeitsplan

Das Projekt ist auf drei Jahre angelegt. Der Arbeitsplan ist in drei Phasen gegliedert, die jeweils ein Jahr dauern:

Jahr 1. Entwurf des probabilistischen Modells, Auswahl von Methoden und Werkzeugen, Prototyp-Implementierung, erste Auswertungsexperimente. Diese Phase ist fast abgeschlossen.

Jahr 2. Tiefere Einsicht in die verwendeten Algorithmen (z.B. Konvergenzanalyse von MCMC-Sampling), Erweiterung des Modells um weitere Merkmale und Kompositionsregeln. Performanzoptimierung und Fehleranalyse.

Jahr 3. Anwendung und Evaluierung des Modells bei praktischen Aufgaben der Sprachtechnologie: OCR-Nachkorrektur, Lemmatisierung, Tagging. Untersuchung zu weiteren Anwendungsmöglichkeiten. Umfassende Evaluierung auf Daten aus verschiedenen Sprachen. Veröffentlichung der Dissertation.

Finanzierung

Die Promotion wird aus den Mitteln des Europäischen Sozialfonds (ESF) der Europäischen Union (EU) und des Freistaates Sachsen im Rahmen des ESF-Projekts Nr. 100234741 „Landesinnovationspromotionen“ finanziert. Projektzeitraum: 01.09.2015–31.08.2018.

Literaturverzeichnis

[AF11] Aronoff, Mark; Fudeman, Kirsten Anne: What is morphology? Fundamentals of Linguistics. Wiley, 2011.

[An92] Anderson, Stephen R.: A-Morphous Morphology. 1992.

[BK03] Beesley, Kenneth R.; Karttunen, Lauri: Finite State Morphology. Center for the Study of Language and Information, 2003.

[BMT02] Baroni, Marco; Matiasek, Johannes; Trost, Harald: Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In: Proceedings of the 6th Workshop of the ACL Special Interest Group on Phonology. Jgg. 6, S. 48–57, 2002.

[Bo08] Bordag, Stefan: Unsupervised and Knowledge-free morpheme segmentation and analysis. Lecture Notes in Computer Science, 5152 LNCS:881–891, 2008.

[Ca11] Can, Burcu: Statistical Models for Unsupervised Learning of Morphology and POS Tagging. Dissertation, University of York, 2011.

[CDvG08] Chrupała, Grzegorz; Dinu, Georgiana; van Genabith, Josef: Learning morphology with morfette. In: Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC '08. S. 2362–2367, 2008.

[Cl02] Clark, Alexander: Memory-Based Learning of Morphology with Stochastic Transducers. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). S. 513–520, 2002.

[CL05] Creutz, Mathias; Lagus, Krista: Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Bericht, 2005.

[DLR77] Dempster, A. P.; Laird, N. M.; Rubin, Donald B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1–38, 1977.

[FSM97] Ford, Alan; Singh, Rajendra; Martohardjono, Gita: Pace Pāṇini: Towards a word-based theory of morphology. American University Studies. Series XIII, Linguistics, Vol. 34. Peter Lang Publishing, Incorporated, 1997.

[Go06] Goldsmith, John: An algorithm for the unsupervised learning of morphology. Natural Language Engineering, 12(1):353, 2006.

[Ha55] Harris, Zellig S: , From phoneme to morpheme, 1955.

[HB11] Hammarström, Harald; Borin, Lars: Unsupervised Learning of Morphology. Computational Linguistics, 37(2):309–350, 2011.

[Ja15] Janicki, Maciej: A Multi-purpose Bayesian Model for Word-Based Morphology. In (Mahlow, Cerstin; Piotrowski, Michael, Hrsg.): Systems and Frameworks for Computational Morphology – Fourth International Workshop, SFCM 2015. Springer, 2015.

[Ki13] Kirschenbaum, Amit: Unsupervised Segmentation for Different Types of Morphological Processes Using Multiple Sequence Alignment. In: 1st International Conference on Statistical Language and Speech Processing, SLSP. Tarragona, Spain, S. 152–163, 2013.

[Ko83] Koskenniemi, Kimmo: Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Dissertation, University of Helsinki, 1983.

[Mi13] Mikolov, Tomas; Corrado, Greg; Chen, Kai; Dean, Jeffrey: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the International Conference on Learning Representations (ICLR 2013). S. 1–12, 2013.

[MYZ13] Mikolov, Tomas; Yih, Wen-tau; Zweig, Geoffrey: Linguistic regularities in continuous space word representations. In: Proceedings of NAACL-HLT. S. 746–751, 2013.

[NF02] Neuvel, Sylvain; Fulop, Sean A.: Unsupervised Learning of Morphology without Morphemes. In: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON). S. 31–40, 2002.

[PCT09] Poon, Hoifung; Cherry, Colin; Toutanova, Kristina: Unsupervised morphological segmentation with log-linear models. June, S. 209, 2009.

[RC05] Robert, Christian P.; Casella, George: Monte Carlo Statistical Methods (Springer Texts in Statistics). Springer-Verlag New York, Inc., 2005.

[Ru13] Ruokolainen, Teemu; Kohonen, Oskar; Virpioja, Sami; Kurimo, Mikko: Supervised Morphological Segmentation in a Low-Resource Learning Setting using Conditional Random Fields. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL). Sofia, Bulgaria, S. 29–37, 2013.

[SCR12] Samdani, Rajhans; Chang, Ming-Wei; Roth, Dan: Unified Expectation Maximization. In: 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. S. 688–698, 2012.

[SO15] Soricut, Radu; Och, Franz Josef: Unsupervised Morphology Induction Using Word Embeddings. In: NAACL 2015. S. 1626–1636, 2015.

[YW00] Yarowsky, David; Wicentowski, Richard: Minimally Supervised Morphological Analysis by Multimodal Alignment. In: ACL '00. S. 207–216, 2000.