Modelle sozialer Netzwerke und Natural Language Processing: eine methodologische Randnotiz

Alexander Mehler[1], Andy Lücking[2]
Goethe Universität Frankfurt, Text Technology Lab

Eine Vielzahl von Modellen sozialer Netzwerke basiert auf der (teil-)automatischen Analyse von Korpora natürlichsprachlicher Texte.[3] Dabei handelt es sich um Korpora, die beispielsweise Daten von Twitter, Facebook, Weblogs, Wikipedia, E-Mail-Systemen oder vergleichbaren Medien umfassen. Diese Art von *Primärdaten* werden – vielfach mit Hilfe von Methoden des *Natural Language Processing* (NLP) – in *Sekundärdaten* (cf. Brinker, Sager 2006) überführt, um hieraus schließlich Netzwerkmodelle von sozialen Systemen als den entsprechenden Modelloriginalen (Stachowiak 1965, 1989) zu gewinnen. Die resultierenden Netzwerkmodelle bilden Daten dritter Ordnung, welche als Input zur Berechnung einschlägiger Netzwerkstatistiken (Newman 2010) dienen (siehe Abbildung 1). Im Vordergrund unserer Notiz zu Netzwerkmodellen stehen solche Verfahren, bei denen Abbildung 2 (von Modellen sprachlicher auf Modelle sozialer Entitäten) mit Methoden des NLP automatisiert durchgeführt wird. Hierzu steht eine Reihe von Werkzeugen bereit, und zwar ausgehend von der so genannten *Tokenisierung* und *Lemmatisierung* über das *Wortarten-Tagging*, die *Named Entity Recognition* (NER) (z.B. von Personen, Orten oder Organisationen), die Erkennung von Zeitausdrücken, die *automatische Disambiguierung*, das *Semantic Role Labelling* (z.B. von *agent*, *patient* und *instrument* einer Handlung) und die *Relation Extraction* bis hin zur *Event Detection*, dem *Topic Tracking* und der *Frame Analysis*, um nur wenige Beispiele zu nennen (siehe Jurafsky, Martin 2000 sowie Manning, Schütze 1999 für Übersichten über diese und verwandte Ansätze). Idealerweise annotieren solche NLP-Methoden sämtliche der in den Inputkorpora manifestierten Informationen derart, dass sie computerbasiert weiterverarbeitet werden können. Dabei sind insbesondere intertextuell konstituierte Informationen relevant, welche dadurch zustande kommen, dass sie Informationen (etwa zu denselben Personen oder denselben Organisationen) aus mehreren Texten aggregieren.

In den resultierenden Netzwerkmodellen der Ebene 3 aus Abbildung 1 denotieren Knoten soziale, situationelle (Barwise, Perry 1983) oder kognitive (Johnson-Laird 1988) Entitäten (z.B. Personen, Organisationen, Institutionen, Orte, Zeiten oder mentale Modelle), während Kanten Prozesse bzw. Relationen (z.B. der Koordination, Kooperation, Kollaboration, der zeitlichen oder räumlichen Inklusion, des kognitiven Alignments – Pickering, Garrod 2004) dieser Entitäten abbilden.

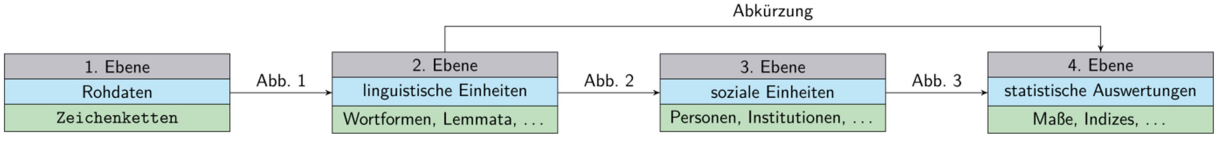

Abbildung 1: Von Textkorpora zu Modellen sozialer Netzwerke und hierauf aufsetzenden Netzwerkstatistiken.

Unser Ansatz besteht nun darin, festzustellen, dass diese kaskadierte Abbildung insbesondere im Hinblick auf den Begriff der *Identität* eine Reihe von methodologischen Problemen aufwirft. Wenn wir beispielsweise einen Knoten *X* eines Netzwerks *N* als Modell einer sozialen Entität *Y* betrachten,

[1] Prof. Dr. Alexander Mehler; mehler@em.uni-frankfurt.de; Fachbereich für Informatik und Mathematik; Goethe-Universität Frankfurt; Robert-Mayer-Straße 10; D-60325 Frankfurt am Main
[2] Dr. Andy Lücking; Luecking@em.uni-frankfurt.de; Fachbereich für Informatik und Mathematik; Goethe-Universität Frankfurt; Robert-Mayer-Straße 10; D-60325 Frankfurt am Main
[3] Dies ist das Ergebnis einer Fragenbogenaktion, welche anlässlich eines interdisziplinären Workshops zu sozialen Netzwerken im Rahmen des Darmstädter Schader-Forums am 25-26.04.2016 durchgeführt wurde.

dann setzen wir im Idealfall voraus, dass sämtliche Informationen über *Y*, die das Inputkorpus *C* bereithält, exploriert wurden, um Knoten *X* in *N* informationell anzureichern bzw. strukturell einzubetten. Jedes Segment von Texten aus Korpus *C* wäre folglich dahingehend zu überprüfen, inwieweit es strukturelle Information dieser Art beinhaltet. Das Problem ist nun, dass diese Aufgabe im Allgemeinen fern davon ist, gelöst zu sein. Mehr noch, ihr Lösungsgrad ist nicht sonderlich gut bekannt – von speziellen Evaluationsszenarien für NLP-Methoden einmal abgesehen, welche jedoch zumeist den „wahren Fehler", wie er aus der Anwendung solcher Methoden resultiert, unterschätzen. Es wäre zumindest nötig, im Vorfeld zu wissen, welche Entitäten überhaupt vernetzt werden sollen, da wir nicht erwarten können, dass ein rein textbasierter Ansatz all diese Informationen einem Textkorpus entnehmen kann. Ein solches ontologisches Modell (Cimiano et al. 2014) liegt der Mehrzahl der statistischen NLP-Ansätze jedoch nicht zugrunde und ist im Allgemeinen nur sehr schwer zu erstellen. An dieser Stelle böte es sich an, und diesen Weg beschreiten offenbar viele Ansätze, nicht etwa Netzwerke von Entitäten der Ebene 3, sondern von Einheiten der Ebene 2 zu betrachten (siehe die „Abkürzung" in Abbildung 1). Hier trifft man jedoch auf dasselbe Problem der Identität, dessen Lösung abermals den Rückgriff zumindest auf eine vorzugebende terminologische Ontologie (Sowa 2000) impliziert. Viele Gattungsnamen sind bekanntermaßen mehrdeutig, so dass man im Zuge der Netzwerkbildung zu disambiguieren hat. Doch welche Bedeutung hat man im konkreten Fall eines Textvorkommens anzusetzen? Im Idealfall klärt uns ein Disambiguierungsmodell wenigstens über die Wahrscheinlichkeitsverteilungen der Lesarten von Wörtern *ex ante* auf – aus den Korpora selbst sind solche Modelle nicht vollständig zu gewinnen, da Mehrdeutigkeit kein rein sprachsystematisch induziertes Problem ist. Modelle, welche Umfang und Verteilung von Lesarten je Wort abschätzen, können anhand von großen Korpora (wie der Wikipedia, welche zudem Disambiguierungsseiten ausweist) gelernt werden. Wir können jedoch nicht sicher voraussagen, dass ein solches Korpus gerade die Mehrdeutigkeitsfälle unseres Modelloriginals abdeckt. An dieser Stelle ließe sich die Fehleranalyse durch Verweis auf die Kontextsensitivität der natürlichen Sprache (Barwise, Perry 1983) und ihre Variationsquellen (Fritz 2006) beliebig fortsetzen. Im Kern stehen wir vor einem Modellierungsproblem, dass mit Fehlerarten von textbasierten Modellen sozialer Netzwerke in Zusammenhang steht:

- **Typ-0-Fehler:** Strings (wie beispielsweise so genannte Boilerplates in Webseiten), die nicht Teil der analyserelevanten Daten sind, werden Korpus *C* zugeschlagen, so dass die strukturelle Einbettung von Knoten letztlich verrauscht wird.
- **Typ-1-Fehler:** Dieselbe Entität (etwa ein Wort (als Modelloriginal von Knoten der Ebene 2) oder eine Person (als Modelloriginal von Knoten der Ebene 3)) wird auf verschiedene Knoten des Netzwerks *N* abgebildet, so dass schließlich auch die Kanten-basierten Repräsentationen ihrer Beziehungen verteilt werden.
- **Typ-2-Fehler:** Derselbe Knoten aus *N* resultiert aus der Aggregation von Informationen zu verschiedenen Entitäten des jeweiligen Modelloriginals. Infolgedessen bildet dieselbe Kante aus *N* möglicherweise verschiedene, zusammenhanglose Prozesse oder Relationen ab.

Mikro-, Meso- oder Makroebenen-bezogene Einheiten, welche aus solchen Netzwerken abgeleitet werden, bergen das Risiko einer Vervielfältigung dieser Fehlerarten auf die jeweilige Ableitungsebene, und zwar so, dass hierauf aufsetzende Statistiken invalide sind. Der Grund hierfür besteht im Kern darin, dass nicht länger von einer Abbildungsbeziehung zwischen (struktureller, semantischer oder funktionaler) Rolle im Modelloriginal und struktureller Position im Netzwerkmodell ausgegangen werden kann. Um Probleme dieser Art anzugehen, benötigen wir Methoden für die Abschätzung von Fehlern der genannten Art. Solche Abschätzungsmethoden

stehen wiederum in Zusammenhang mit Sensitivitätsanalysen, welche bei zu variierender Genauigkeit und Konzertierung der eingesetzten NLP-Methoden Abschätzungen darüber geben, wie sich die entsprechenden Fehlerraten verändern. Solche Sensitivitätsanalysen fehlen in dem hier untersuchten Bereich nahezu vollständig. Ganz unabhängig von dieser Einschätzung stellen wir in Abrede, dass NLP-Methoden quasi aus Textkorpora allein valide Modelle sozialer Netzwerke unüberwacht lernen können. Hierfür bedarf es vielmehr einer modelltheoretischen Semantik des jeweiligen Modelloriginals, welche im Bereich rein statistischer NLP-Methoden noch immer eine untergeordnete Rolle spielen.

Literatur

Barwise, J., Perry, J. 1983. Situations and Attitudes: Cambridge: MIT Press.

Brinker, K., Sager, S.F. 2006: Linguistische Gesprächsanalyse. Berlin: Erich Schmidt Verlag.

Cimiano, P., Unger, C., McCrae J. 2014: Ontology-based interpretation of natural language. Toronto: Morgan & Claypool Publishers.

Fritz, G. 2006: Historische Semantik. Stuttgart: J. B. Metzler'sche Verlagsbuchhandlung.

Jurafsky, D., Martin, J.H. 2000: Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Upper Saddle River: Prentice Hall.

Manning, C.D., Schütze, H. 1999: Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: MIT Press.

Newman, M.E.J. 2010: Networks: An Introduction. Oxford: Oxford University Press.

Sowa, J.F. 2000: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Pacific Grove: Brooks/Cole.

Stachowiak, H. 1965: Gedanken zu einer allgemeinen Modelltheorie. In: Studium Generale 18.7, S. 432–463.

Stachowiak, H. 1989: Modell. In H. Seiffert, G. Radnitzky (Hg), Handlexikon zur Wissenschaftstheorie. München: Ehrenwirth, p. 219–222.