Exposé zur Studienarbeit

Simultanes Lernen von Wortarten und Eigennamen durch diskriminatives Training eines Graphischen Modells

Fabian Mößner

Machine Learning Group TU-Berlin

Student der Humboldt-Universität zu Berlin

Betreuer: Dr. Ulf Brefeld

1 Einleitung

Ein überwältigend großer Teil der heute verfügbaren Informationen ist weltweit in Texten natürlicher Sprache enthalten. Es ist immer noch eine grosse Herausforderung die in diesen Daten enthaltenen Informationen auch automatisiert zugänglich zu machen, indem die zugrundeliegenden Gesetzmäßigkeiten maschinell gelernt werden. Wichtige Schritte im Gesamtprozess der Maschinellen Sprachverarbeitung (NLP) sind die Wortartenzuordnung (POS) und die Eigennamenerkennung (NER). Bei letzterer wird den Wörtern zugeordnet, ob sie einer Kategorie wie Ortsname, Organisation oder ähnlichem entsprechen. Momentan lernen die meisten Verfahren die verschiedenen Teilprozesse des NLP einzelnd und riskieren so ein Aufsummieren von Fehlern. In dieser Arbeit werden die Eigennamen gemeinsam mit den Wortarten gelernt, um so durch die Einbeziehung ihrer Abhängigkeiten bessere und robustere Lernergebnisse zu erzielen.

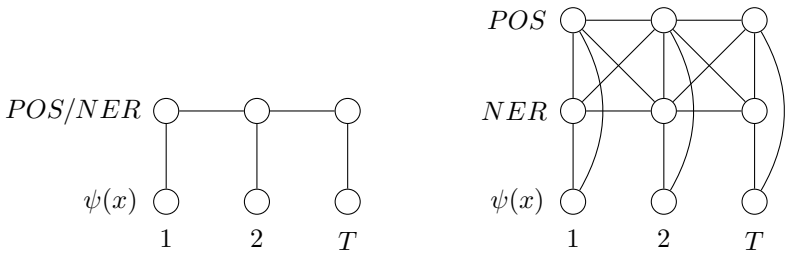
2 Verwandte Arbeiten

Aus dem Maschinellem Lernen kommen zur NER verschiedene Ansätze. Eine wichtige Rolle spielen dabei Graphische Modelle, deren Parameter auf generative [Rabiner 1989], aber auch auf diskriminative Weise gelernt werden können. Die letztere Herangehensweise, wie Conditional Random Fields (CRF) [Lafferty 2001] und Hidden Markov Support-Vector-Machines (HMSVM) [Altun 2003, Tsochantaridis 2005, Altun 2006], sind aufwändiger zu errechnen, erzielen aber empirisch bessere Ergebnisse.
Aber gerade bei den spezifischen Problemen der NER verspricht nicht nur eine sequentielle Modellierung Vorteile, sondern auch das zusätzliche simultane Lernen (Multitask Learning) [Collobert 2008, Argyriou 2006] von mehreren Prozessen der Sprachverarbeitung. Denn es existiert eine prinzipiell unbegrenzt grosse Menge an Eigennamen, die in Bereichen wie der Biomedizin auch ständig erweitert wird. Zusätzlich können mehrere Wörter einen einzelnen Eigennamen bilden. Deshalb liegt das Einbeziehen von weiterem Wissen und Abhängigkeiten von anderen Worteigenschaften hier besonders nahe.

3 Vorhaben

In dieser Arbeit wird nun das mit einem Perzeptron trainierte CRF [Altun 2003] so erweitert, dass ein Multitask-Lernen von POS und NER möglich ist. Das Modell auf dem aufgebaut und mit dem verglichen werden soll ist in Abbildung (1a) zu sehen. Während hier nur NER- bzw. POS-Label gelernt werden, werden in dem erweiterten

Graphischem Modell (Abbildung (1b)) die beiden Label gemeinsam in einem einzigen Prozess gelernt. Hier fließen nun auch die Äbhängigkeiten der beiden Label zueinander ein. Während beim ersten Modell Viterbi als Inferenzalgorithmus eingesetzt werden kann, ist dies bei dem erweiterten komplexerem Modell nicht möglich. Stattdessen müssen allgemeinere Inferenzverfahren wie z. B. Loopy Belief Propagation eingesetzt werden. Die Implementation der Graphischen Modelle, sowie deren Inferenz werden in dieser Arbeit mithilfe der Bayes Net Toolbox von Kevin Murphy realisiert [BNT] . Der Korpus für die empirische Evaluation bildet die Nachrichtensammlung RCV1 der Agentur Reuters. Dieser ist vollständig mit POS- und NER-Tags annotiert, wobei die Eigennamenkategorien zwischen beginnenden und fortgesetzten Orts-, Organisations-, Personen- und sonstigen Namen unterscheiden. Mit der Kategorie für Wörter, die keine Eigennamen sind, ergibt sich so eine Gesamtzahl von neun Klassen für die NER. Die Feature, die die Wörter in dem Korpus repräsentieren wurden bereits an der TU-Berlin generiert. Gemessen werden soll die durchschnittliche Fehlerrate pro Sequenz und das Laufzeitverhalten der Verfahren in Relation zur Größe der Trainingsdaten. Als ein Erfolg des Vorhabens wäre eine um 0.05 verbesserte Fehlerrate bei der NER zu werten.



(a) Single-Perzeptron (b) Multitask-Perzeptron

Abbildung 1: Die Graphischen Modelle für einen Satz mit T Wörtern. Ein Knoten in der Zeile $\psi(x)$ steht für die Featurerepräsentation eines Wortes. In (a) sind die Abhängigkeiten von $\psi(x)$ zu POS bzw. NER-Label und vom Label y_t zu y_{t+1} modelliert. In (b) kommen die Abhängigkeiten der beiden verschiedenen Labelsequenzen hinzu.

Literatur

[Tsochantaridis 2005] I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun: *Large Margin Methods for Structured and Interdependent Output Variables*. Journal of Machine Learning Research 6, 2005

[Altun 2003] Y. Altun, I. Tsochantaridis, T. Hofmann: *Hidden Markov Support Vector Machines*. ICML, 2003

[Altun 2006] Y. Altun, T. Hofmann, I. Tsochantaridis: *SVM Learning for Interdependent and Structured Output Spaces*. 2006

[Rabiner 1989] Lawrence R. Rabiner: *A tutorial on Hidden Markov Models and selected applications in speech recognition*. Proceedings of the IEEE 77, 1989

[Collins 2002] M. Collins: *Discriminative training methods for Hidden Markov Models: Theory and experiments with Perceptron Algorithms*. Proceedings of the Conference on Empirical Methods in NLP, 2002

[Lafferty 2001] John Lafferty, Andrew McCallum, Fernando Pereira: *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Int. Conf. Machine Learning, 2001

[Borthwick 1998] A. Borthwick, J. Sterling, E. Agichtein, R. Grishman: *Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition*. WVLC, 1998

[Collobert 2008] R. Collobert, J. Weston: *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning*. 25th ICML, 2008

[Argyriou 2006] A. Argyriou, T. Evgeniou, M. Pontil: *Multi-Task Feature Learning*. NIPS, 2006

[BNT] Kevin Murphy: *Bayes Net Toolbox*. http://www.apastyle.org/elecref.html, 2002