Seminar Natural Language Processing

Lehrstuhl für Grundlagen der Programmierung

Wintersemester 2020/21

Die meisten Paper stammen aus dem Gebiet des Konstituentenparsings. Dabei wird einem gegebenen Satz aus einer natürlichen Sprache ein *Konstituentenbaum* zugewiesen (siehe Figure 1 links). Dieser zeigt, wie der Satz in seine syntaktischen Bestandteile (*Konstituenten*) zerlegt (*geparst*) werden kann. Im Beispiel wird etwa der gesamte Satz (S) in eine Nominalphrase (NP), die das Wort *John* umspannt, und eine Verbalphrase (VP), die die Wörter *loves Mary* umspannt, zerlegt. Konstituentenparsing kann z.B. durch kontextfreie Grammatiken (CFG) modelliert werden, wobei die Nichtterminale den syntaktischen Kategorien und die Terminale den Wörtern der natürlichen Sprache entsprechen. Die Regeln spiegeln die Zerlegung der Konstituenten wider, so wird beispielsweise die oben geschilderte Zerlegung durch die Regel S $\rightarrow$ NP VP modelliert. Mithilfe von Algorithmen wie dem CYK-Algorithmus können für gegebene Sätze Ableitungen der CFG berechnet werden. Aus diesen werden dann die Konstituentenbäume abgelesen.

Im obigen Beispiel war jeder Konstituent *kontinuierlich*, d.h. er hat einen zusammenhängenden Teil des Satzes überspannt. Im Gegensatz dazu überspannen *diskontinuierliche Konstituenten* mehrere, nicht zusammenhängende Teile eines Satzes (siehe Figure 1 rechts). Sie können nicht durch CFG modelliert werden und sind aktiver Forschungsgegenstand.

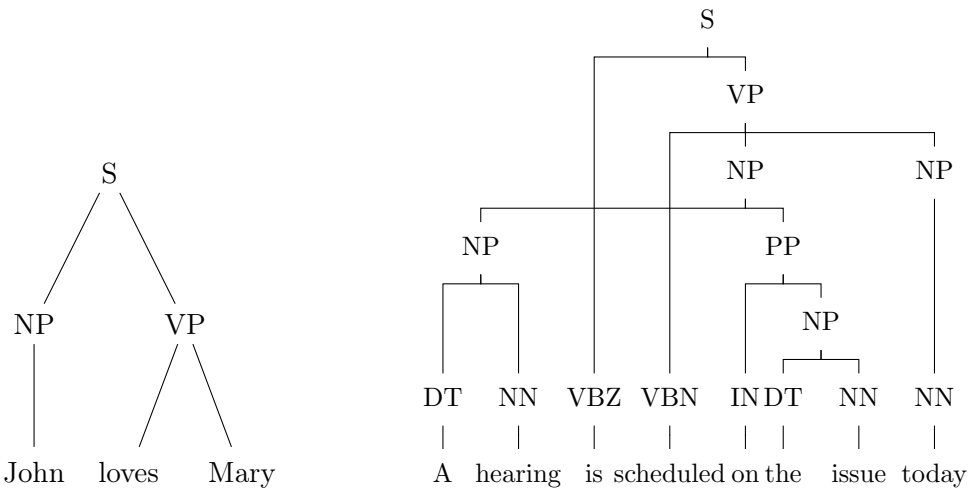

Figure 1: Links: Konstituentenbaum für *John loves Mary*. Rechts: diskontinuierlicher Konstituentenbaum für *A hearing is scheduled on the issue today*.

Themen für das Proseminar

Tetra-Tagging: Word-Synchronous Parsing with Linear-Time Inference (Kitaev and Klein 2020). Das Paper präsentiert einen Ansatz für kontinuierliches Konstituentenparsing, der ohne eine kontextfreie Grammatik auskommt. Mithilfe eines neuronalen Netzes wird für jede Position im Eingabesatz vorhergesagt, wie diese in den Parsebaum einzuordnen ist. Anschließend bestimmt ein Algorithmus mittels dynamischer Programmierung aus diesen Vorhersagen in linearer Zeit einen Parsebaum.

Span-based discontinuous constituency parsing: a family of exact chart-based algorithms with time complexities from $O(n^6)$ down to $O(n^3)$ (Corro 2020). Das Paper präsentiert einen Ansatz für diskontinuierliches Konstituentenparsing ohne explizites Grammatikmodell. Der Parser arbeitet auf Grundlage eines gewichteten Deduktionssystems, wobei die Berechnung der Gewichte durch ein neuronales Netz erfolgt. Dabei wird experimentell untersucht, wie sich die Einschränkung der Komplexität auf Parse-Zeit und Parse-Genauigkeit auswirkt.

Discontinuous Constituency Parsing with a Stack-Free Transition System and a Dynamic Oracle (Coavoux and Cohen 2019). Das Paper präsentiert einen Ansatz für diskontinuierliches Konsituentenparsing mit Transitionssystemen. Neu ist dabei, dass als Datenstruktur eine Menge statt wie üblich ein Stack verwendet wird, wodurch Parsing in linearer Zeit möglich wird. Die Wahl der Transitionen erfolgt mithilfe eines neuronalen Netzes.

Discontinuous Constituent Parsing as Sequence Labeling (Vilares and Gómez-Rodríguez 2020). Das Paper reduziert diskontinuierliches Konstituentenparsing auf *Sequence Labeling*, also die Zuweisung eines Labels zu jeder Position im Eingabesatz. Die Auswahl der Labels wird von einem neuronalen Netz getroffen.

Conditional random fields: Probabilistic models for segmenting and labeling sequence data (Lafferty, McCallum, and Pereira 2001). Das Paper führt *Conditional Random Fields (CRF)*, ein diskriminatives Modell, ein und demonstriert dessen Einsatz im Sequence Labeling. Es zeigt durch theoretische Argumente und Experimente die Überlegenheit von CRF gegenüber vorherigen (insbesondere generativen) Modellen wie z.B. Hidden Markov Models.

Towards Probabilistic Acceptors and Transducers for Feature Structures (Quernheim and Knight 2012). Das Paper führt gewichtete Automaten und Transducer für gerichtete azyklische Graphen (directed acyclic graph; DAG) ein. Diese stellen eine Erweiterung der geläufigeren Automaten- und Transducermodelle für Strings und Bäume dar und können zur Analyse der Bedeutung eines Satzes eingesetzt werden. Das Paper skizziert außerdem deduktive Algorithmen zur Lösung des Membership-Problems und zur Berechnung der besten Ableitung.

Themen für das Hauptseminar

Span-Based LCFRS-2 Parsing (Stanojević and Steedman 2020). Das Paper präsentiert einen Ansatz für diskontinuierliches Konstituentenparsing. Er baut auf einer Verallgemeinerung des CYK-Algorithmus für LCFRS-2 auf, ohne jedoch explizit eine Grammatik zu modellieren. Die Gewichtsberechnung erfolgt durch neuronale Netze.

The Problem with Probabilistic DAG Automata for Semantic Graphs (Vasiljeva, Gilroy, and Lopez 2019). Das Paper untersucht, inwiefern gewichtete Automaten für gerichtete azyklische Graphen (DAG-Automaten) Wahrscheinlichkeitsverteilungen modellieren können. Es wird gezeigt, dass es für bestimmte Typen von DAG-Automaten durch Zuweisung von Gewichten zu Transitionen lediglich triviale Wahrscheinlichkeitsverteilungen modelliert werden können.

References

Coavoux, Maximin and Shay B. Cohen (June 2019). "Discontinuous Constituency Parsing with a Stack-Free Transition System and a Dynamic Oracle". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 204–217. DOI: 10.18653/v1/N19-1018. URL: https://www.aclweb.org/anthology/N19-1018.

Corro, Caio (2020). *Span-based discontinuous constituency parsing: a family of exact chart-based algorithms with time complexities from $O(n^6)$ down to $O(n^3)$*. arXiv: 2003.13785 [cs.CL]. URL: https://arxiv.org/abs/2003.13785.

Kitaev, Nikita and Dan Klein (July 2020). "Tetra-Tagging: Word-Synchronous Parsing with Linear-Time Inference". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6255–6261. DOI: 10.18653/v1/2020.acl-main.557. URL: https://www.aclweb.org/anthology/2020.acl-main.557.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In: URL: https://repository.upenn.edu/cis_papers/159/.

Quernheim, Daniel and Kevin Knight (July 2012). "Towards Probabilistic Acceptors and Transducers for Feature Structures". In: *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Jeju, Republic of Korea: Association for Computational Linguistics, pp. 76–85. URL: https://www.aclweb.org/anthology/W12-4209.

Stanojević, Miloš and Mark Steedman (July 2020). "Span-Based LCFRS-2 Parsing". In: *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*. Online:

Association for Computational Linguistics, pp. 111–121. DOI: 10.18653/v1/2020.iwpt-1.12. URL: https://www.aclweb.org/anthology/2020.iwpt-1.12.

Vasiljeva, Ieva, Sorcha Gilroy, and Adam Lopez (June 2019). "The problem with probabilistic DAG automata for semantic graphs". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 902–911. DOI: 10.18653/v1/N19-1096. URL: https://www.aclweb.org/anthology/N19-1096.

Vilares, David and Carlos Gómez-Rodríguez (2020). *Discontinuous Constituent Parsing as Sequence Labeling*. arXiv: 2010.00633 [cs.CL]. URL: https://arxiv.org/abs/2010.00633.