

SWP 22/23

Projektvorschläge Steen

Applied Timeline Summarization

- Timeline-Summarization erzeugt aus Dokumenten eine strukturierte Timeline-Darstellung von Ereignissen
- Ereignisse werden einem Datum zugeordnet und zusammengefasst
- Ziel: Zeitliche Übersicht über ein Ereignis erlangen

2011-03-15
First protests after calls on Facebook for a “Day of Dignity.”
2011-08-18
US President Barack Obama and his allies urge Assad to quit. Western and Arab states later impose sanctions on his regime.
2011-10-02
Creation of the opposition Syrian National Council SNC.

Beispieltimeline

Projektidee

- TLS wird auf fixen Datensätzen aus Dokumenten mit Gold-Timelines evaluiert
- Systeme sind i.d.R. relativ unflexibel implementiert
- **Projektziel:** Setze ein TLS-System in die Praxis um

Projektziele

- Reimplementiere einen TLS-Ansatz und evaluiere auf Standard-Datensätzen
- Integriere Ansatz in eine nutzbare Oberfläche für TLS
 - Nutzer gibt Query (=Thema der Timeline), Zeitraum und Länge an
 - System erzeugt Timeline basierend auf Angaben
- Mögliche Datenquellen:
 - Feststehendes Corpus (z.B. kleines Subset von Common Crawl, gigaword)
 - Wikipedia Current Events Portal
 - Search-API + Eigenes Crawling

Herausforderungen

- Auswahl eines passenden Algorithmus für Anforderungen
- Reimplementation unter Rücksichtnahme auf die verschiedenen Einsatzszenarien (Evaluation/Validierung auf Standarddaten vs. Einsatz im System)
- Umgang mit voraussichtlichen Usability-Problemen (z.B. lange Zeiten für Generierung)
- Identifikation von Performanzverbesserungen (Parallelisierbarkeit etc.)

TLS-Algorithmen

- TLS wird meistens unsupervised gelöst
- **Martschat and Markert (2018):** globale Optimierung mit submodularen Funktionen basierend auf Satzähnlichkeiten und Datumshäufigkeiten
- **Ghalandari and Ifrim (2020)** geben eine gute Übersicht über Ansätze
 - Datumsweise Zusammenfassung
 - Cluster-basierte Zusammenfassung auf Event Ebene

Literatur

- Ghalandari, Demian Gholipour, and Georgiana Ifrim. "Examining the State-of-the-Art in News Timeline Summarization." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.
- Martschat, Sebastian, and Katja Markert. "A Temporally Sensitive Submodularity Framework for Timeline Summarization." Proceedings of the 22nd Conference on Computational Natural Language Learning. 2018.

Heideltime Reimplementierung

- Heideltime (Strötgen und Gertz, 2012) ist ein Time Expression Tagger
- Findet beispielsweise für TLS Verwendung
- Verwendung mit Python ist aber umständlich und integriert nicht gut mit Python Workflow
- Projektziel: Reimplementierung von Heideltime in Python/Python-Extension

Heideltime

- Heideltime ist ein regelbasiertes System
- Muster identifizieren Ausdrücke, zweites Muster regularisiert
- Bestimmte Muster haben zusätzlich Part-of-Speech Restriktionen
- Beispiel für February 25, 2009
- `EXTRACTION=„(%reMonthLong|%reMonthShort)(%reDayNumberTh|
%reDayNumber)[\s]?,? %reYear4Digit(, %reWeekday)?“`
- `NORM_VALUE="group(7)-%normMonth(group(1))-
%normDay(group(4))"`

Projektziele

- Reimplementierung von Heidelttime mit Originalressourcen (s. <https://github.com/HeidelTime/heidelttime>) in Python Modul
- Implementierung entweder in reinem Python, oder als Python Extension
- Kooperation mit TLS Projekt möglich, aber nicht notwendig

Literatur

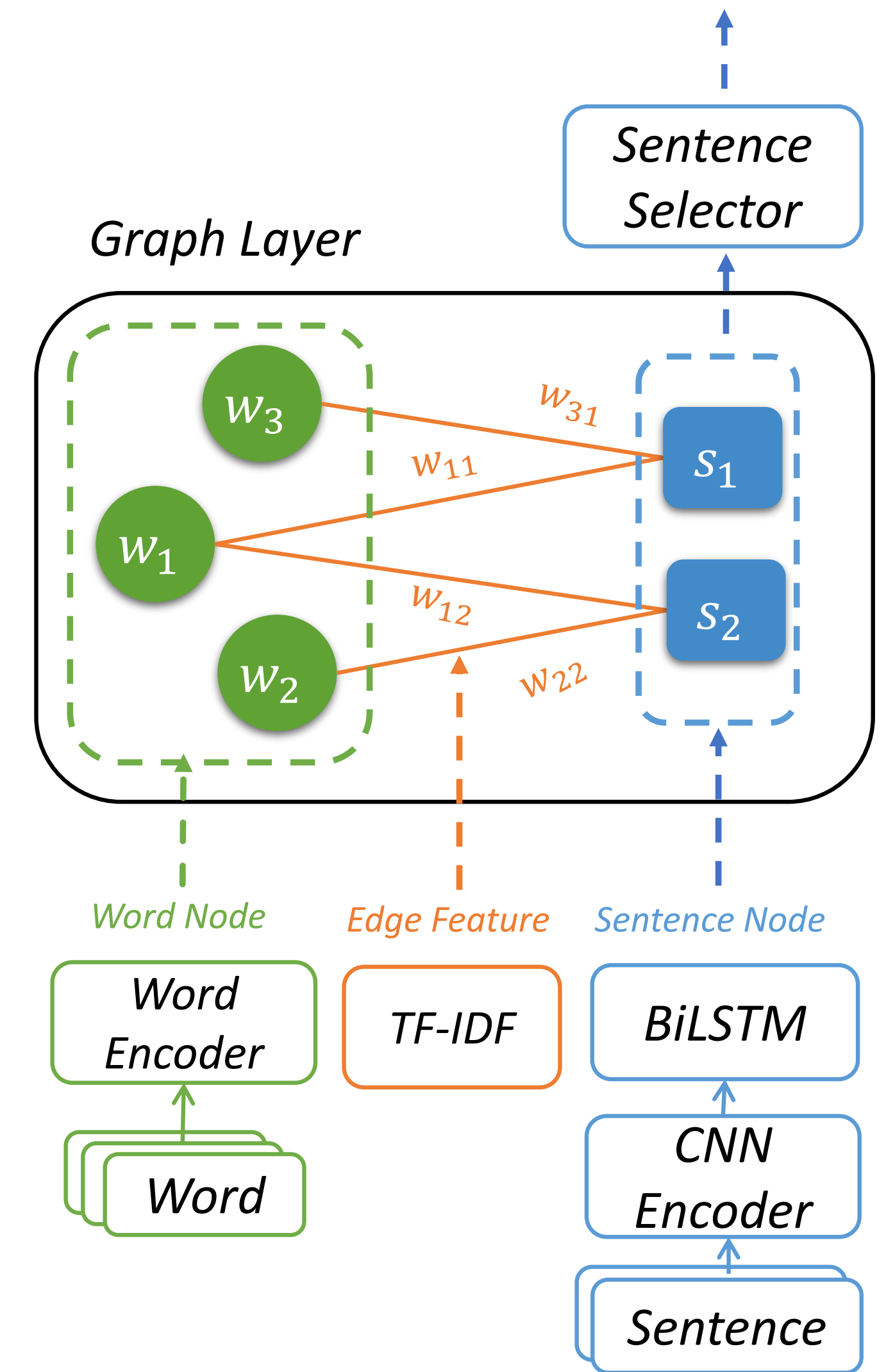
- Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Extractive Summarization with Graphs

- Extraktive Summarization identifiziert relevante Sätze in einem Dokument als Zusammenfassung
- Viele Summarization Systeme repräsentieren Dokumente über Graphstrukturen (Radev, 2004, Mihalcea and Tarau, 2004),
- Modernerer Ansatz: Graphstruktur + NN = GNN
- Projektidee: Reimplementierung/Erweiterung eines Graph-basierten Summarization Systems

Heterogenous Graph NNs

- Eine Option: Heterogenous Graph Network (Wang et al., 2020)
- Wort/Satz Graph mit Graph Attention Network
- Mögliche Erweiterungen:
 - Verbesserte Kantengewichte
 - Topicmodelling



Literatur

- Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous Graph Neural Networks for Extractive Document Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6209–6219, Online. Association for Computational Linguistics.
- Baoyu Jing, Zeyu You, Tao Yang, Wei Fan, and Hanghang Tong. 2021. Multiplex Graph Neural Network for Extractive Text Summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 133–139, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.