Multiword Expressions

WS 2019/20

Institut für Computerlinguistik, Heidelberg

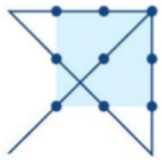
Mark-Christoph Müller

Natural Language Processing Group

Heidelberg Institute for Theoretical Studies gGmbH

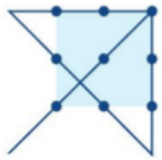
mark-christoph.mueller@h-its.org

Organisatorisches



- Kurssprache? Referat / HA auch auf Englisch möglich!
- Bitte keine Handys, Tablets, Notebooks etc.
- In Liste eintragen
- Überblick
  - Semester / Fächerkombi / Vorkenntnisse?
  - Wer will welchen Schein?
- Sprechstunde (bei Bedarf):
  - INF 327, SR 4, vor oder direkt nach dem Kurs (dann kürzer)
  - Wenn **vor** dem Kurs, bitte per Mail anmelden!
- Technik für Präsentationen

Scheinerwerb I: Mitarbeit 1

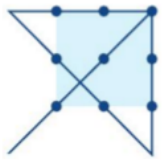


- Anwesenheit
  - max. zwei mal unentschuldigtes Fehlen (Anwesenheitsliste)
- Aktive Beteiligung an der Diskussion
- Behandelte Literatur
  - vorab lesen
  - mind. eine **Frage** / Papier (lt. Plan) formulieren und bis Montag vor Kurs, 12:00 an

    `mark-christoph.mueller@h-its.org`

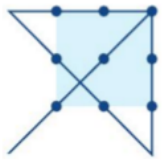
    mit Subject: MWE <Datum der Sitzung>
- **Erstmalig für Sitzung am 05.11.2019**

Scheinerwerb I: Mitarbeit 2



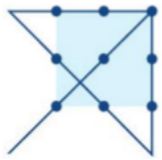
- **Fragen**
  - Wurde die Literatur gelesen?
  - Kritische Auseinandersetzung, z.B.
    - “Warum machen die Autoren ABC?”
    - “Woher kommt XYZ?”
  - Begründen!
  - Verständnisfragen (Ausnahme!)
- Als eigene Vorbereitung auf Sitzung, und für Diskussion
- Für mich als Feedback
- **Gewichtung Mitarbeit: 30%**

Scheinerwerb II: Referat 1



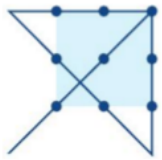
- Referat über ein bis zwei Papiere (werden vorgegeben)
  - zusätzliche Literatur nach Wunsch / Rücksprache
- Unterscheidung nach PS / HS
- Ggfs. auch zwei Personen pro Referat (dann etwas höhere Anforderungen, Anteile müssen erkennbar sein)
- 60 min + anschließende Diskussionsleitung
- **Besprechung jeweils eine Woche vorher!**

Scheinerwerb II: Referat 2



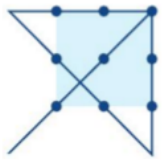
- **Inhalt (60% der Referatsnote)**
  - Identifikation des zentralen Beitrags / der zentralen Aussage
  - Beschreibung der technischen Lösung und der Evaluation
  - Berücksichtigung der wichtigsten Sekundär-Literatur
  - Eigene Bewertung:
    - Stärken und Schwächen des Papiers
    - Verbesserungspotential
  - Methoden-Exkurs

Scheinerwerb II: Referat 3



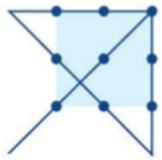
- **Präsentation (20% der Referatsnote)**
  - Aufbau und Qualität der Folien
  - Struktur des Vortrags
  - Klarheit
  - Beispiele
  - weiteres Hintergrundmaterial
  - Verständlichkeit des mündlichen Vortrags

Scheinerwerb II: Referat 4



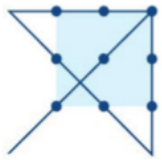
- **Diskussionsleitung (20% der Referatsnote)**
  - Wird eine Diskussion angeregt?
  - Können Fragen beantwortet werden?
- **Gewichtung Referat: 40%**
- Je nach Teilnehmerzahl: Zweitreferat (statt HA / Projekt) möglich, bitte nachfragen!

Scheinerwerb III: Hausarbeit



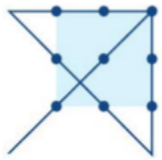
- **Ausarbeitung des Referats**
  - Einordnung in größeren Forschungszusammenhang
  - Beurteilt wird u.a.
    - Struktur, Klarheit
    - Hinzunahme von / Vergleich mit anderen Papieren / Ansätzen
    - Eigene Bewertung
    - Formales: Zitationsformate, Gliederung
  - 14 / 10 Seiten max. (HS / PS) plus Literatur, 11pt, 1.5-zeilig, PDF
  - **Rücksprache vor Beginn!**
  - **Abgabetermin: 14. April 2020, 17:00**
- **Gewichtung Hausarbeit: 40%**

Scheinerwerb III: Programmier-Projekt



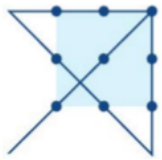
- **Implementierung** eines existierenden / neuen eigenen Ansatzes
  - Achtung Umfang: Schaffbar in ca. 6 Wochen?
  - Je nach Thema: Mögliche Grundlage einer BA- oder MA-Arbeit
  - **Projektbeschreibung** (hauptsächl. Bewertungsgrundlage)
    - Forschungsplan
    - Umsetzung / Dokumentation
    - Evaluation / Reproduzierbarkeit
    - Präsentation / **Demo**
  - 7 Seiten max. 11pt, 1.5-zeilig, PDF
  - **Rücksprache vor Beginn!**
  - **Abgabetermin: 14. April 2020, 17:00**
- **Gewichtung Programmier-Projekt: 40%**

Tipps fürs Referat I



- **Einstieg**
  - Kritisches, aufmerksames Lesen der Papiere und der wichtigsten Referenzen
- **Verständnisprobleme?**
  - Zitierte Literatur und Sekundärliteratur lesen
  - ***Danach*** Frage aufbereiten und in Vorbesprechung ansprechen
- Stellen Sie das Papier/die Papiere kritisch im wissenschaftlichen Kontext dar!
- Achten Sie auf einen roten Faden!
- Nicht “Related Work” referieren!
- Konzentrieren Sie sich auf entscheidende Punkte!

Tipps fürs Referat II

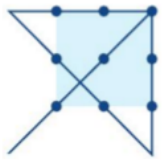


- **Leitfragen**
  - Mit welchen Fragestellungen befassen sich die Papiere?
  - Wie hängen diese mit den Fragestellungen des Seminars zusammen?
  - Inwiefern unterscheiden sich die in den Papieren beschriebenen Ansätze untereinander (und von vorhergehender Literatur)?
  - Was sind die Innovationen (theoretisch, methodisch, etc.)?
  - Gibt es Probleme bei der Evaluierung? Sind die Resultate überzeugend?
  - Wie haben Ihnen die Papiere gefallen?

Tipps fürs Referat III

- **Durchführung**
  - Nicht von den Folien ablesen!
  - Die Zuhörer anschauen (und nicht die Folien)!
  - Nicht erst ***während des Vortrags*** nach Formulierungen oder Beispielen (!) suchen!
  - Schwierige / wichtige Teile explizit vorformulieren (und dann natürlich frei vortragen!)
  - Üben! Zeit stoppen!

Wie geht’s weiter?



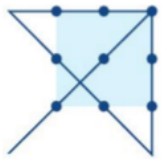
- Fragen?
- Kurzvorstellung der Referatsthemen
- Erstes Referat am **05.11.2019** (wird heute vergeben)
- Weitere Referate **ab 12.11.2019**
  - Mail mit **drei** Wunschthemen (Datum/Referenzen) an mich bis 28.10.2019, 12:00, mit Subject: **MWE Referat**
  - Info: Bachelor mit % / Master, PS / HS, Interesse Zweitreferat?
- **29.10.2019**
  - Vor dem Seminar: Besprechung erstes Referat
  - Bekanntgabe der Einteilung
  - Einführung durch mich

MWE *Discovery* (Constant et al. 2017) I

- Erstellung eines Inventars von *potentiellen* MWE-Typen (*types*)

- Basierend auf **statistischer** Analyse, hier:
  - MWE = Sequenz von “Einheiten” mit größerer als erwarteter Häufigkeit (=Kollokation)
  - Kein Rückgriff auf Semantik, daher
    - geringe Präzision (viele uninteressante Sequenzen)
    - geringer Recall (seltene, interessante Sequenzen fehlen)
- **Papiere**
  - Church & Hanks (1990): PMI
  - Dunning (1993): log-likelihood
- **Sitzung am 05.11.2019**

MWE *Discovery* (Constant et al. 2017) II



- Erstellung eines Inventars von *potentiellen* MWE-Typen (*types*)

- Basierend auf **statistischer** Analyse, hier:
  - MWE = Gruppe von “Einheiten” innerhalb eines Wort-Fensters (beliebige Länge)
- **Papiere**
  - Smadja (1993), **pp. 150-166**: AVG und STDDEV der Distanzen
  - Colson (2017): CPR-Score (auch multi-lingual)
- **Sitzung am 12.11.2019**

MWE *Discovery* (Constant et al. 2017) III

- Erstellung eines Inventars von *potentiellen* MWE-Typen (*types*)

- Basierend auf **semantischer** Analyse, hier:
  - `sem([AB]) ≠ sem([A]) + sem([B])`
- **Papiere**
  - Zhai (1997): BOW-basierter Vergleich von Kontexten
  - Salehi et al. (2015): Vektor-basiert, Englisch & Deutsch
- **Sitzung am 19.11.2019**

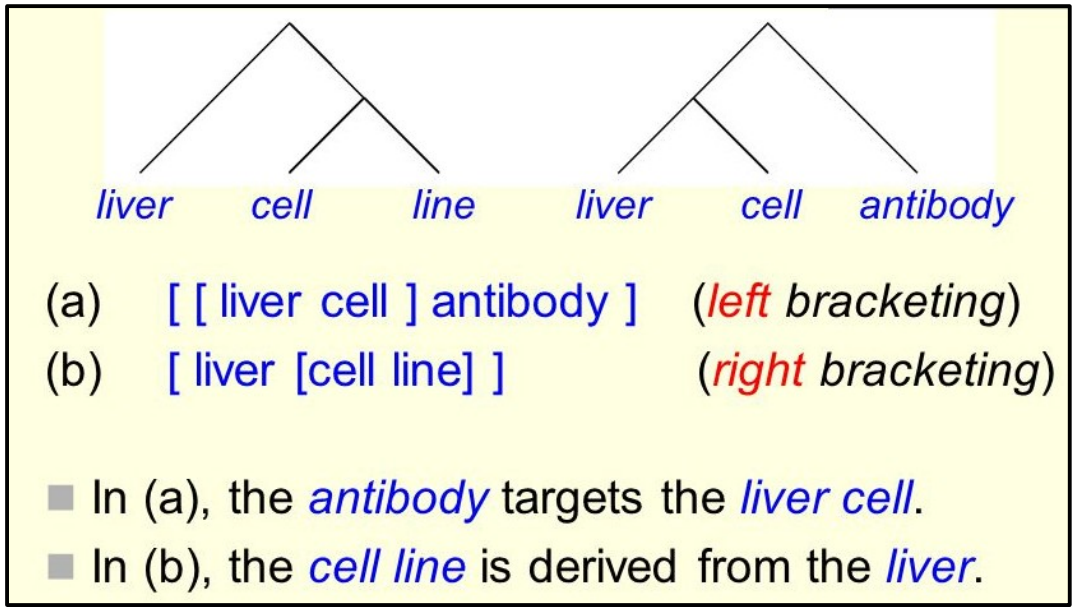
MWE *Identification* (Constant et al. 2017)

- Erkennung von MWE-Instanzen (*tokens*) im Text

- Nicht alle Vorkommen von *potentiellen* MWEs sind *tatsächlich* welche: ‘ins Wasser fallen’ ; ‘make a face’
- **Papiere**
  - Katz & Giesbrecht (2006): LSA, Deutsch
  - Cook et al. (2007): Syntaktische Regularitäten
- **Sitzung am 26.11.2019**

MWE *Bracketing*

- Erkennung der internen Struktur von MWEs mit >2 Bestandteilen
- Meistens: Noun-Noun compounds



Nakov & Hearst (2005)

- **Papiere**
  - Nakov & Hearst (2005), Pitler et al. (2010): Web-Corpus-basiert
  - Barrière (2014): Wikipedia-basiert
- **Sitzung am 03.12.2019**

Noun-Noun Compound Analysis I

- Interpretation der semantischen **Relation**

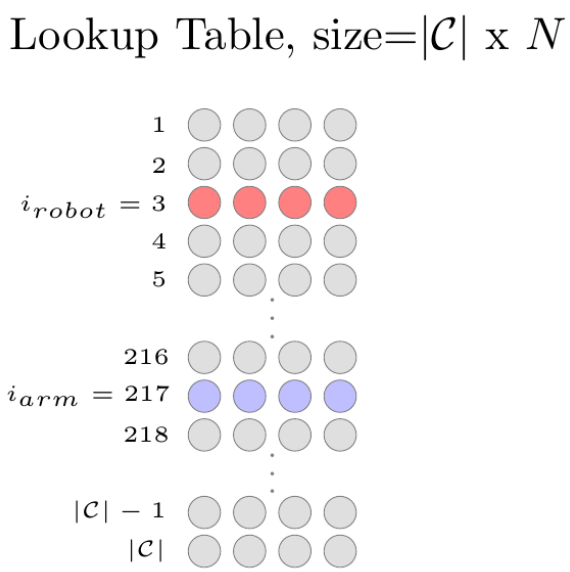
N1 CAUSE N2	sex scandal, withdrawal symptom
N2 CAUSE N1	tear gas, shock news
N1 HAVE N2	lemon peel, school gate
N2 HAVE N1	camera phone, picture book,
N1 MAKE N2	court order, snowball
N2 MAKE N1	computer industry, silk worm
N2 USE N1	steam iron, wind farm
N2 BE N1	island state, soldier ant
N2 IN N1	field mouse, letter bomb
N2 FOR N1	arms budget, steak knife
N2 FROM N1	business profit, olive oil
N2 ABOUT N1	tax law, love letter

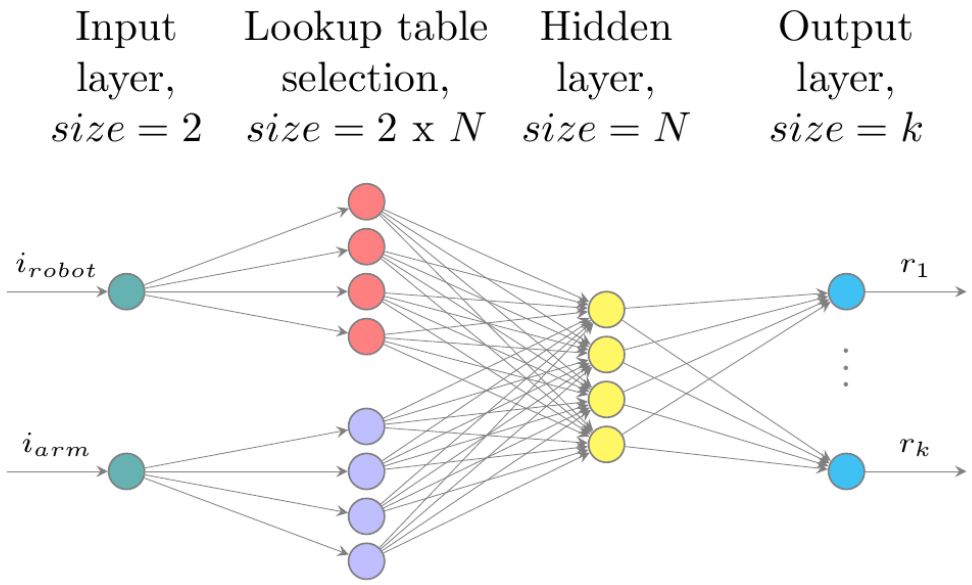
(Bauer & Tarasova 2010)

- **Papiere**
  - Nastase & Szpakowicz (2003): WordNet und Roget's
  - Butnariu & Veale (2008): Web-Corpus-Suche von **Paraphrasen**
- **Sitzung am 10.12.2019**

Noun-Noun Compound Analysis II

- Interpretation der semantischen **Relation**
- Hier: **Machine Learning (Word Embeddings + Neuronales Netz)**

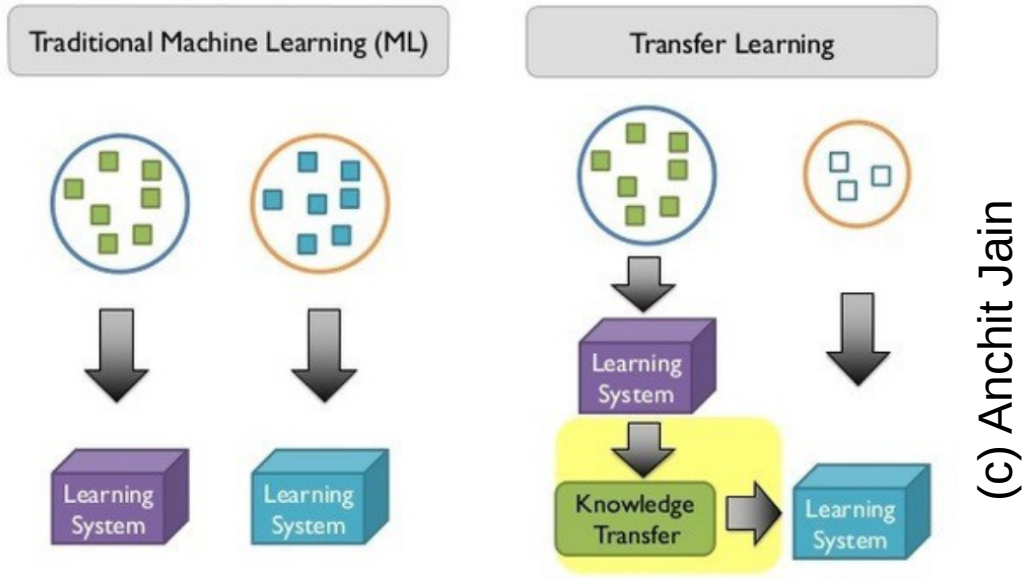




- **Papiere**
  - Dima & Hinrichs (2015), Dima (2016)
- **Sitzung am 17.12.2019**

Noun-Noun Compound Analysis III

- Interpretation der semantischen **Relation**

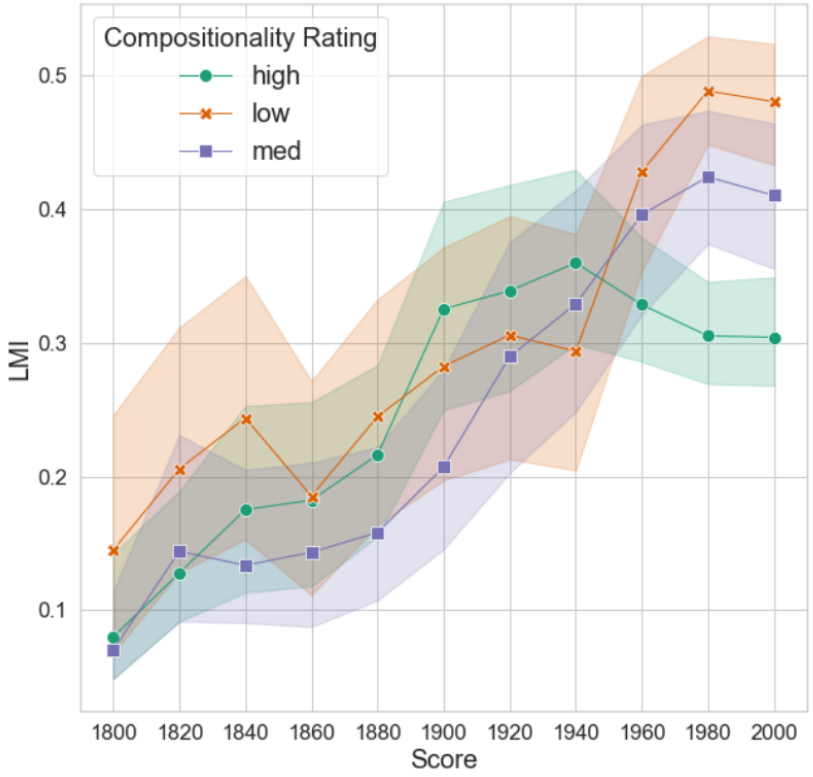


- Hier: durch **Deep Learning**

  - **Papiere**
    - Fares et al. (2018): Transfer und Multi-Task Learning
  - **Sitzung am 07.01.2020**

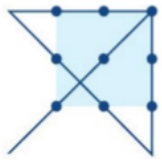
Noun-Noun Compound Analysis IV

- **Historischer Wandel** von Noun-Noun Compounds



- **Papiere**
  - Reddy et al. (2011): Datensatz + Modellierung
  - Dhar et al. (2019): Web-Corpus (diachron)
- **Sitzung am 14.01.2020**

MWEs → Embeddings → NLP-Input

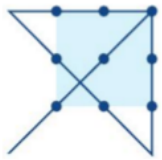


- **Papiere**
  - Mikolov (2013): word2vec + naive ngrams (inkl. **`GoogleNews-vectors-negative300.bin.gz`**)
  - Legrand & Collobert (2016): explizites Lernen von Phrasen
  - Zhao et al. (2017): ngram2vec
- **Sitzung am 21.01.2020**

MWEs in Anwendungen

- **Papiere**
  - Kim et al. (2018): MWEs in Biomedizin (PubMed)
  - Acosta et al. (2011): Information Retrieval
  - Ghoneim & Diab (2013): Machine Translation
- **Sitzung am 28.01.2020**

Abschlusssitzung



- *Lessons learned* etc.
- Hausarbeiten / Programmierprojekte
-
- **Sitzung am 04.02.2020**

Fragen?