Semantic Relation Extraction and Classification in Scientific Papers

Abschlusspräsentation

Florian Abt, Kristina Tesch

20. September 2018

Gliederung

1. Aufgabenstellung (Auffrischung)

2. Stand Zwischenpräsentation

3. Was ist neu?

3.1 Kombination von Klassifikatoren

4. Ergebnisse für die Testdaten

Aufgabenstellung (Auffrischung)

Datenformat (SemEval 2018, Task 7)

- Abstracts von wissenschaftlichen Veröffentlichungen
- Relationen zwischen Entitäten

Beispiel:

```
<text id="H05-2007">
  <title>Pattern Visualization for Machine Translation
       Output</title>
  <abstract>  We describe a method for identifying
      systematic <entity id="H05-2007.1">patterns</
      entity> in <entity id="H05-2007.2">translation
      data</entity> using <entity id="H05-2007.3">part
      -of-speech tag sequences</entity>.  ...
  </abstract>
</text>
```

```
PART_WHOLE(H05-2007.1,H05-2007.2)
```

Datenformat (SemEval 2018, Task 7)

- Abstracts von wissenschaftlichen Veröffentlichungen
- Relationen zwischen Entitäten

Beispiel:

```
<text id="H05-2007">
  <title>Pattern Visualization for Machine Translation
       Output</title>
  <abstract>  We describe a method for identifying
      systematic <entity id="H05-2007.1">patterns</
      entity> in <entity id="H05-2007.2">translation
      data</entity> using <entity id="H05-2007.3">part
      -of-speech tag sequences</entity>.  ...
  </abstract>
</text>
```

```
PART_WHOLE(H05-2007.1,H05-2007.2)
```

Datenformat (SemEval 2018, Task 7)

- Abstracts von wissenschaftlichen Veröffentlichungen
- Relationen zwischen Entitäten

Beispiel:

```
<text id="H05-2007">
  <title>Pattern Visualization for Machine Translation
       Output</title>
  <abstract>  We describe a method for identifying
      systematic <entity id="H05-2007.1">patterns</
      entity> in <entity id="H05-2007.2">translation
      data</entity> using <entity id="H05-2007.3">part
      -of-speech tag sequences</entity>.  ...
  </abstract>
</text>
```

```
PART_WHOLE(H05-2007.1,H05-2007.2)
```

Datenformat (SemEval 2018, Task 7)

- Abstracts von wissenschaftlichen Veröffentlichungen
- Relationen zwischen Entitäten

Beispiel:

```
<text id="H05-2007">
  <title>Pattern Visualization for Machine Translation
       Output</title>
  <abstract>  We describe a method for identifying
      systematic <entity id="H05-2007.1">patterns</
      entity> in <entity id="H05-2007.2">translation
      data</entity> using <entity id="H05-2007.3">part
      -of-speech tag sequences</entity>.  ...
  </abstract>
</text>
```

```
PART_WHOLE(H05-2007.1,H05-2007.2)
```

Datenformat (SemEval 2018, Task 7)

- Abstracts von wissenschaftlichen Veröffentlichungen
- Relationen zwischen Entitäten

Beispiel:

```
<text id="H05-2007">
  <title>Pattern Visualization for Machine Translation
       Output</title>
  <abstract>  We describe a method for identifying
      systematic <entity id="H05-2007.1">patterns</
      entity> in <entity id="H05-2007.2">translation
      data</entity> using <entity id="H05-2007.3">part
      -of-speech tag sequences</entity>.  ...
  </abstract>
</text>
```

```
PART_WHOLE(H05-2007.1,H05-2007.2)
```

Aufgabenstellung (SemEval 2018, Task 7)

Aufgabenteil 1 (Klassifikation)

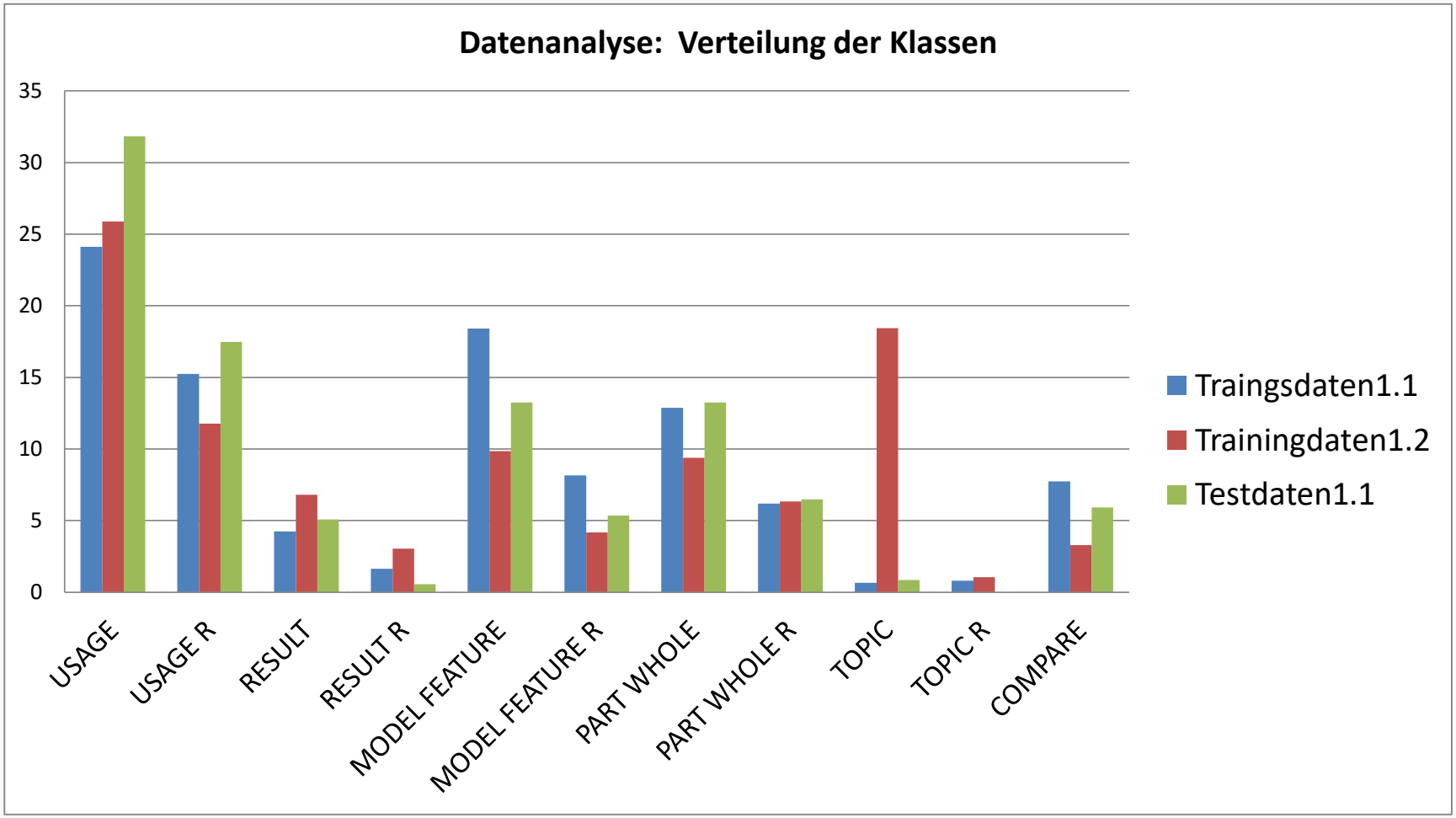
- gegeben: IDs von zwei Entitäten
- gesucht: Relation der Entitäten
  - Usage, Result, Model-Feature, Part-Whole, Topic, Compare
  - die Richtung der Beziehung (nicht für Compare)
- zwei Versionen: manuell annotierte Entitäten und automatisch annotierte Entitäten

Datenanalyse

Aufgabe	1.1	1.2	Test 1.1
#Abstracts	350	350	150
#Relationen	1228	1248	355
#Entitäten	5259	11911	2245

- Sind die Entitäten einer Relation stets in einem Satz?
- Wie verteilen sich die Relationen über die Klassen?
- Gibt es Häufungen in den Entitäten und stehen diese im Zusammenhang mit der Relationsklasse?
- Treten Worte (z.B. than, achieve) häufig in Kombination mit einer Klasse auf?

Klassenaufteilung



Stand Zwischenpräsentation

Bisherige Lösungsansätze

Baseline Classifier

- Auswahl der häufigsten Klasse

$\Rightarrow$ 26,42% und 0,038 F1

Bisherige Lösungsansätze

Baseline Classifier

- Auswahl der häufigsten Klasse

$\Rightarrow$ 26,42% und 0,038 F1

- Klassifikation mit SVM
  - LightRel (Wort-Cluster, Kontext als One-Hot) [RN18]
  - Wortsequenzen (Satzteile, ganzer Satz)
  - Word-Embeddings (Word2Vec, ELMo)
  - Verknüpfen von Entitäten
  - Reverse von Wordsequenzen
  - Vorkommen von häufigen Worten einer Wortart (POS-Feature)
- Klassifikation mit neuronalen Netzen

Bisherige Lösungsansätze

Baseline Classifier

- Auswahl der häufigsten Klasse

$\Rightarrow$ 26,42% und 0,038 F1

- Klassifikation mit SVM
  - LightRel (Wort-Cluster, Kontext als One-Hot) [RN18]
  - Wortsequenzen (Satzteile, ganzer Satz)
  - Word-Embeddings (Word2Vec, ELMo)
  - Verknüpfen von Entitäten
  - Reverse von Wordsequenzen
  - Vorkommen von häufigen Worten einer Wortart (POS-Feature)
- Klassifikation mit neuronalen Netzen

$\Rightarrow$ Bestes Ergebnis 64,96% mit 0,58 F1

Was ist neu?

Features

- SVM:
  - Normalisierung der Wörter
  - Wortreihenfolge nach POS
- CNN:
  - Fehler ist behoben
- Kombination von Classifiern:
  - Mehrheitsentscheidung
  - Einbeziehung nach Klassenwahrscheinlichkeit
  - Mehrheitsentscheid mit Klassenwahrscheinlichkeitsausschluss

Wort:	Anzahl
words	133
word	74
Word	14
Words	6
system	220
systems	49
Systems	13
System	11
corpus	133
Corpus	18

Weshalb kombinieren?

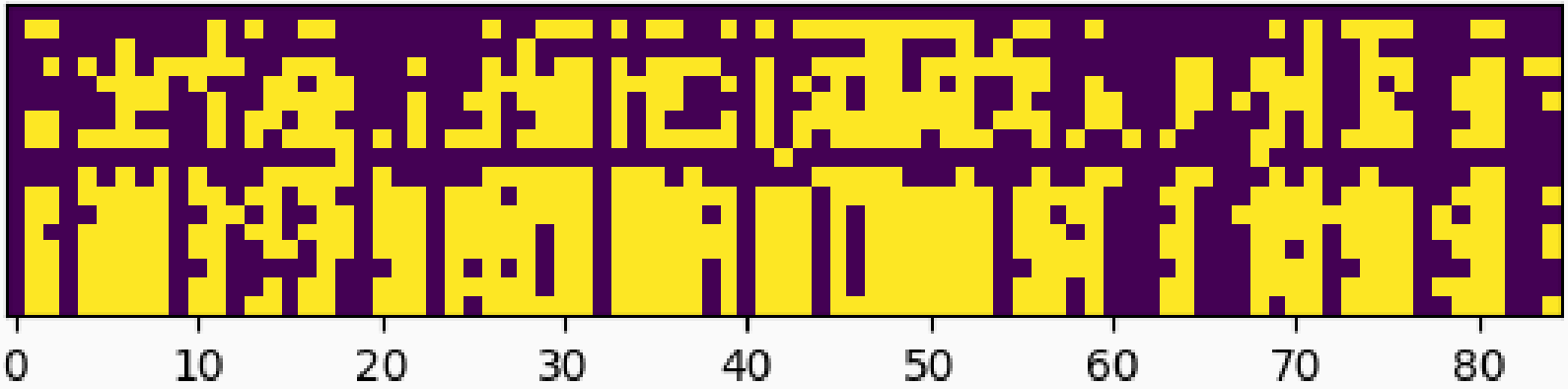


Abbildung 1: Ergebnisse der Basisklassifikatoren für eine Klasse (binär)

Mehrheitsentscheidung I

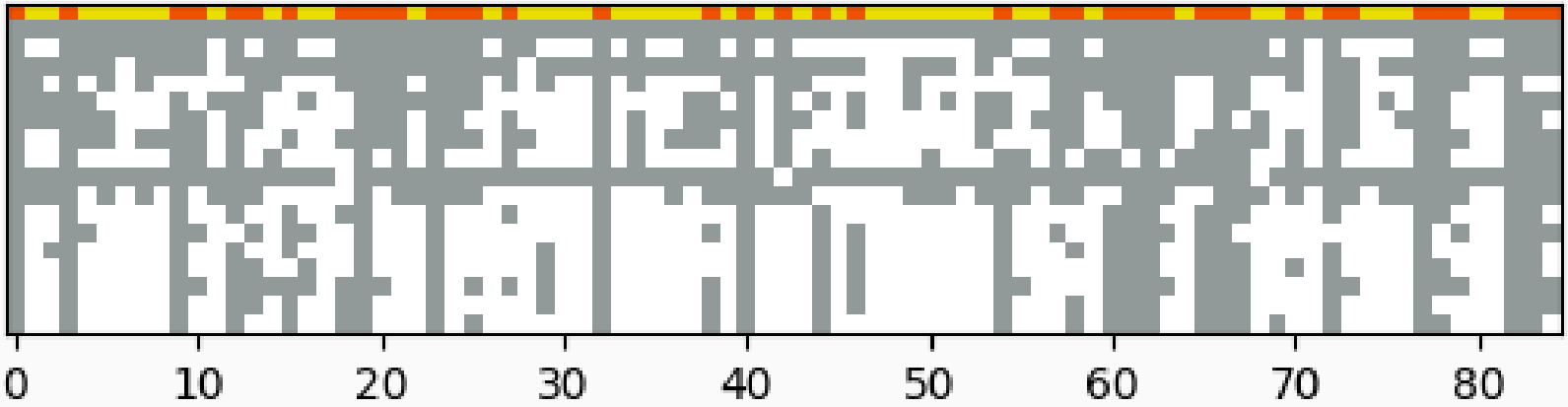


Abbildung 2: Ergebnisse des Mehrheitsvotums für eine Klasse

Mehrheitsentscheidung II

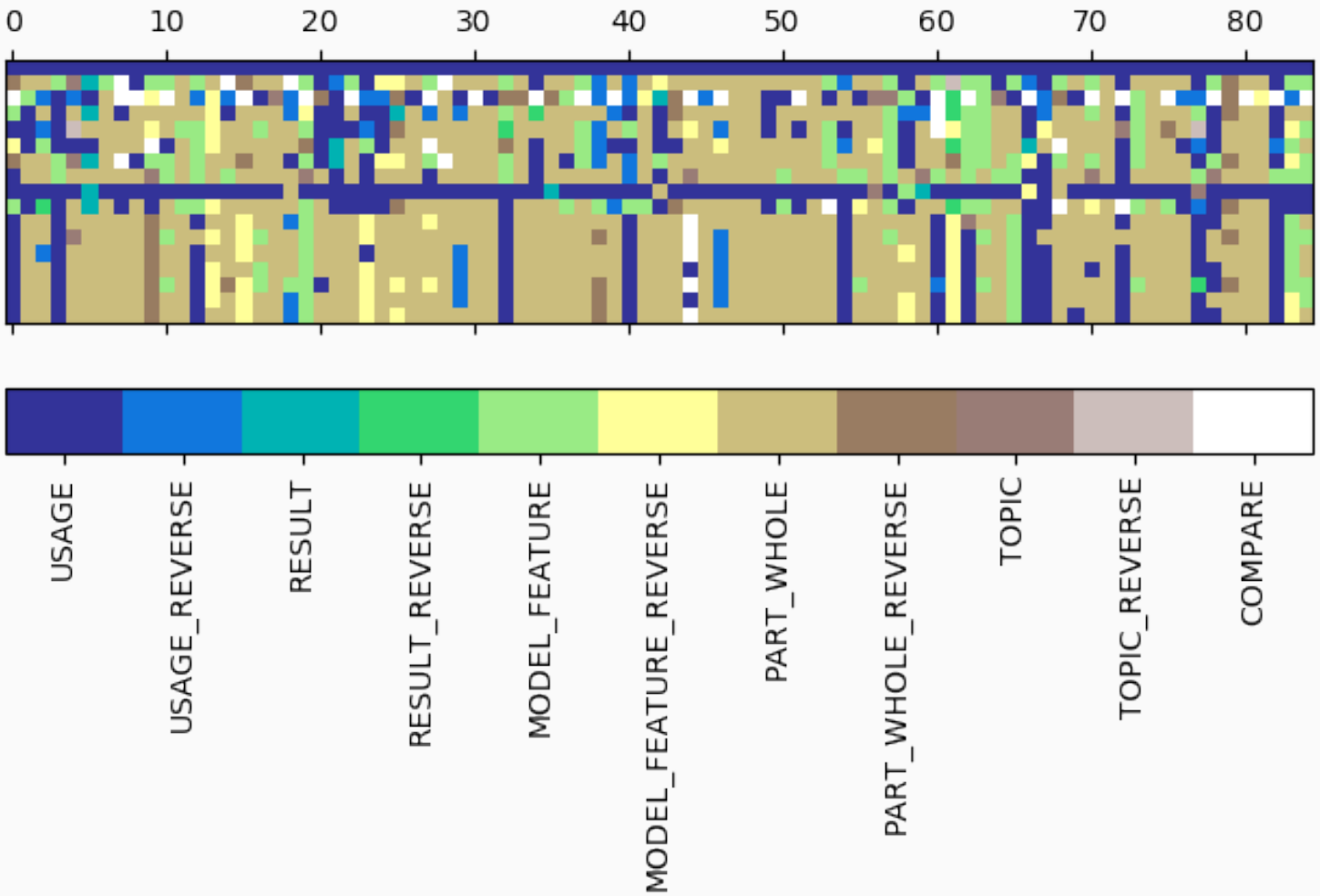


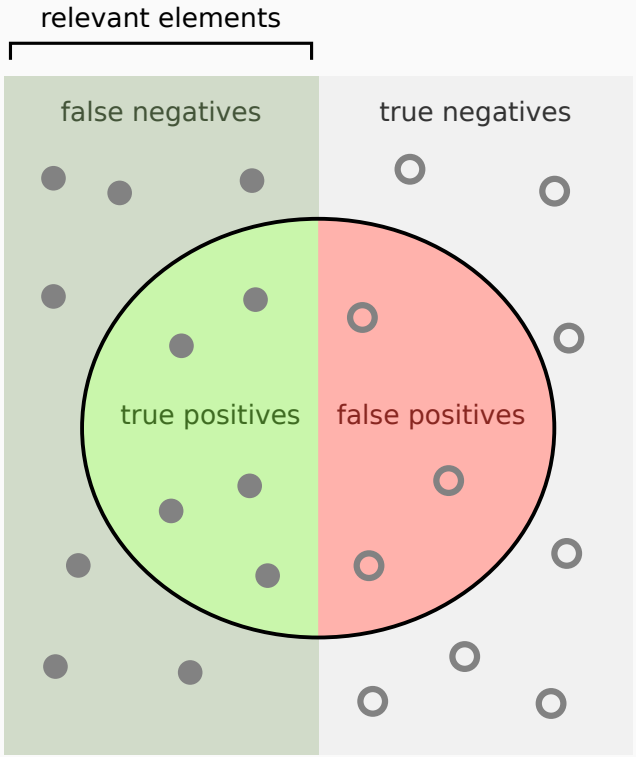
Abbildung 3: Ergebnisse der Basisklassifikatoren für eine Klasse

Mehrheitsentscheidung III

Auflösung eines Gleichstandes:

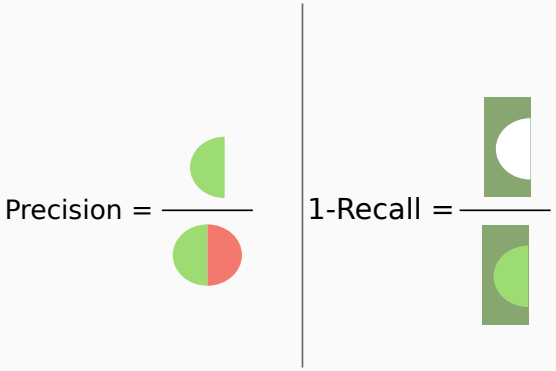
- geschätzte Klassenwahrscheinlichkeiten
- geschätze bedingte Klassenwahrscheinlichkeiten

Mehrheitsentscheidung mit Bedingung



Bedingung pro Classifier pro Klasse:

- Precision $> 0{,}3$
- 1-Recall $< 0{,}9$

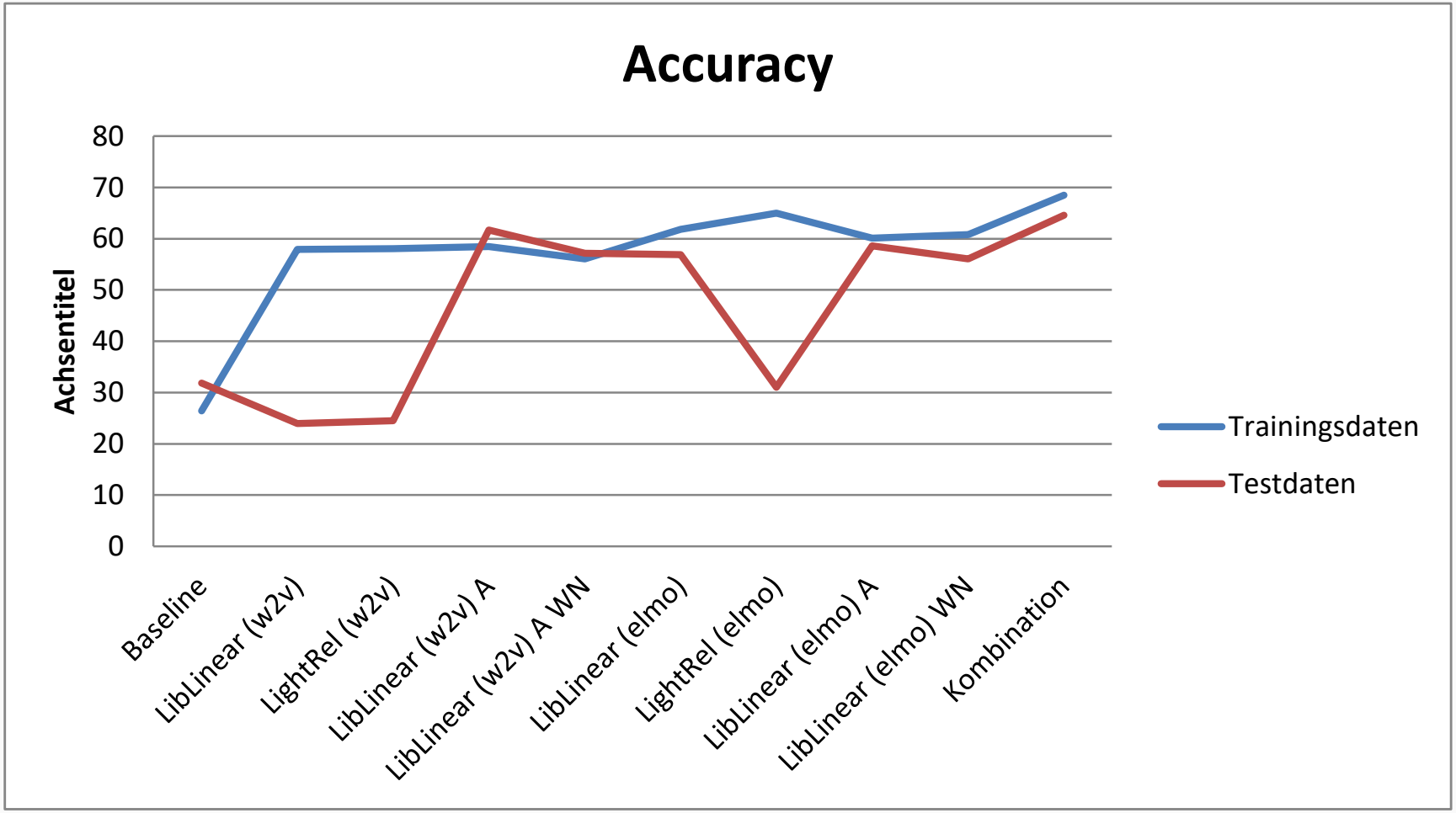


Weitere Ansätze für die Kombination

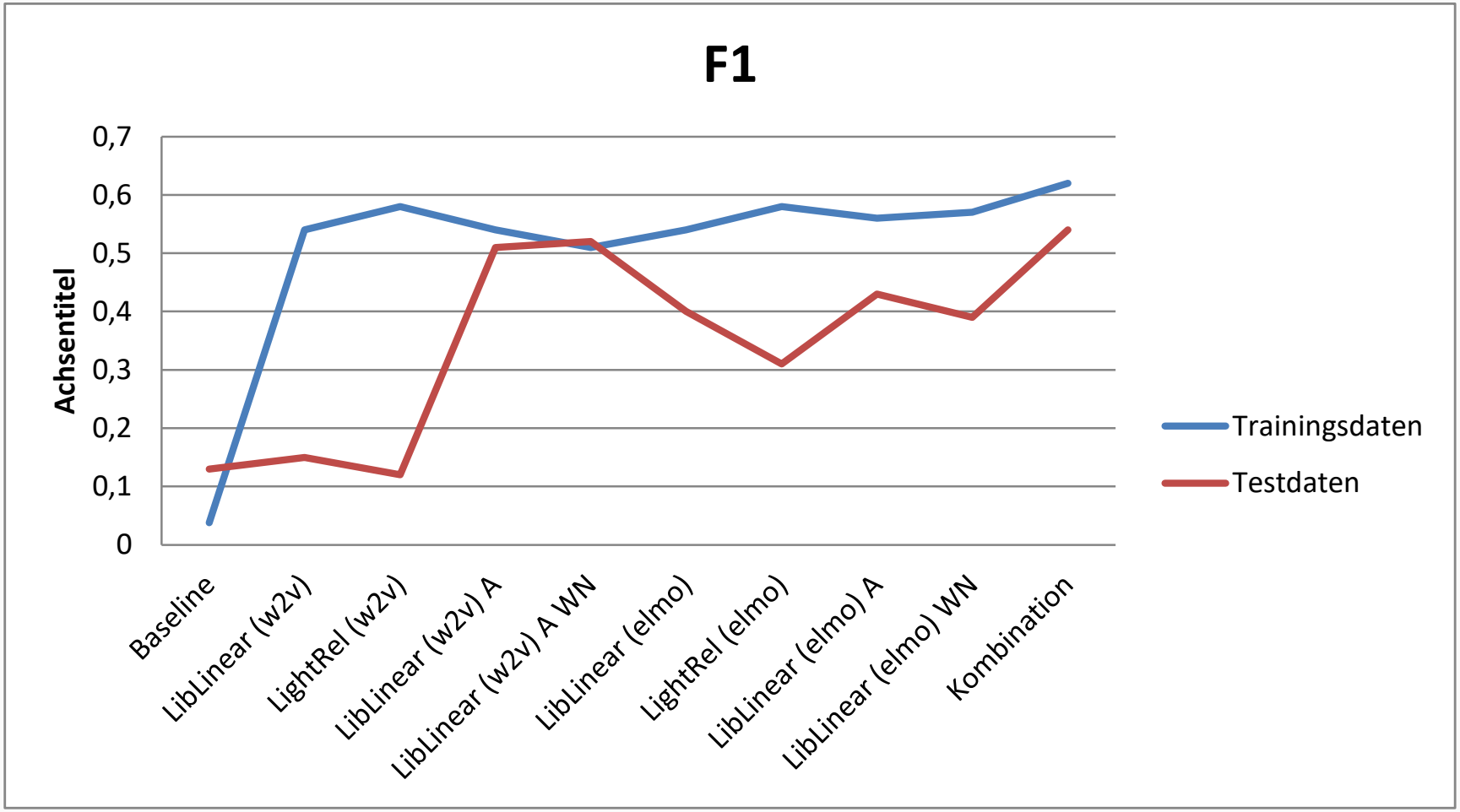
- Einfluss basierend auf den geschätzten bedingten Wahrscheinlichkeiten (wähle Maximum)
- Einfluss SVM nach verschiedenen Features (z. B. Sequenzlänge, Satzlänge, Vorkommen von Wortarten)

Ergebnisse für die Testdaten

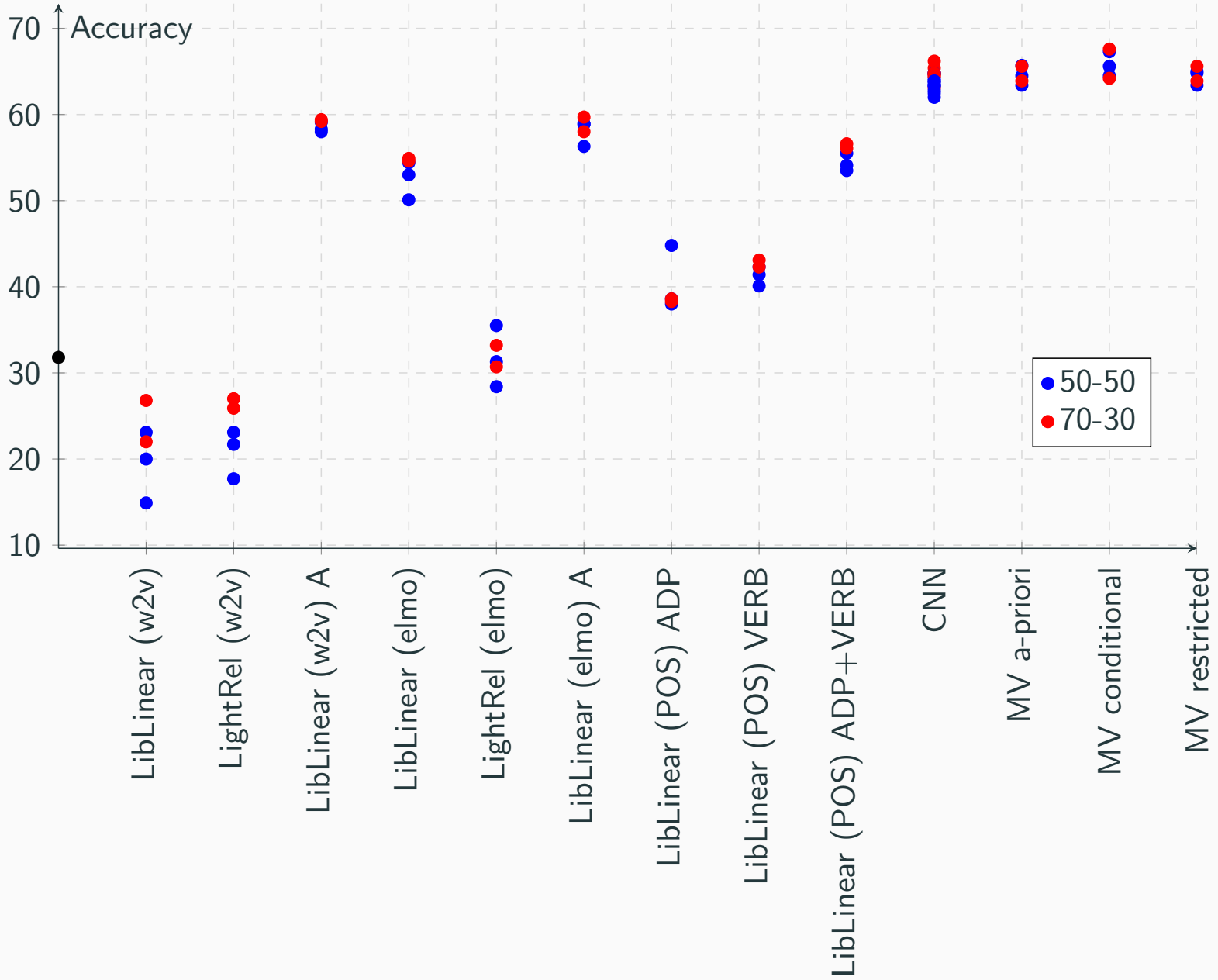
Finale Ergebnisse Accuracy



Finale Ergebnisse F1



Finale Ergebnisse bei verschiedenen Datenaufteilungen



Literatur

Literatur

Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang.
ETH-DS3Lab at SemEval-2018 Task 7: Effectively Combining Recurrent and Convolutional Neural Networks for Relation Classification and Extraction.
CoRR, abs/1804.02042, 2018.

T. Renslow and G. Neumann.
LightRel SemEval-2018 Task 7: Lightweight and Fast Relation Classification.
ArXiv e-prints, April 2018.