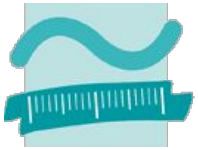


BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

Named Entity Recognition auf Basis von Wortlisten

Lukas Abegg - Humboldt Universität zu Berlin

Tom Schilling - Beuth Hochschule für Technik Berlin

Abschlusspräsentation

SS 2017

> Problemstellung

> Umsetzung

> Evaluation

> Ausblick

> Problemstellung

> Umsetzung

> Evaluation

> Ausblick

Am Anfang war eine Frage...

Welche Ergebnisse können erzielt werden, wenn Tasty zusätzlich auf Basis von Wortlisten funktioniert?

Wie funktioniert Tasty aktuell?

> Einführung Tasty

Tasty arbeitet mit Named Entity Recognition (NER)

- zur Identifikation und Klassifizierung von Named Entities
- vordefinierte Kategorien (Personen, Organisationen oder Orte...)

- NER trainiert gegen Texte (Golden Records)
- durch Named Entity Linking (NEL) wird die Identität von Entitäten bestimmt

Sprachliche Mehrdeutigkeiten werden durch Entity Disambiguation aufgelöst:

> "***Paris*** ist die Hauptstadt von Frankreich"
> => ***Paris*** verweist auf die ***Stadt Paris*** und nicht die ***Person Paris Hilton***

Für das Training wird viel Kontext und dadurch viele Trainingsdaten benötigt

Also was ist die Problemstellung?

Entitäten erkennen, die sich *nicht*
durch den *Kontext erschließen* lassen.

Und dazu der Lösungsansatz?

Die *Named Entity Recognition* auf Basis von *Named Entity Listen* trainieren.

Wie wurde vorgegangen?

Zunächst mussten Entitätsklassen gefunden werden, die Tasty nur schwer erkennen kann.
=> Krankheiten und deren Eigenschaften (Symptome, Wirkstoffe, Behandlungen)

Und wieso genau diese?

Zu Krankheiten *fehlen Gold-Standards,*
die zum Trainieren *genügend groß* sind!

> Problemstellung

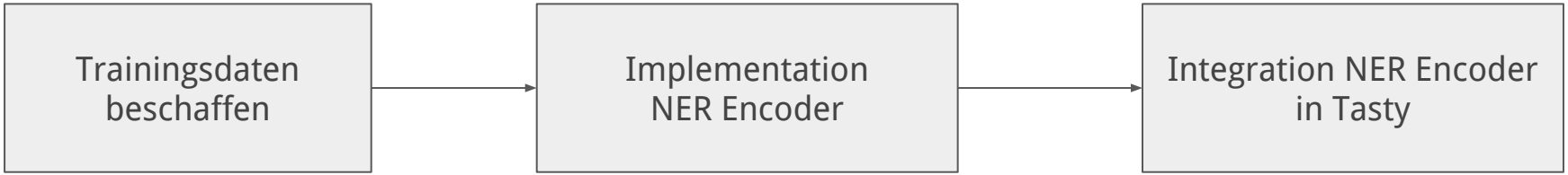
> Umsetzung

> Evaluation

> Ausblick

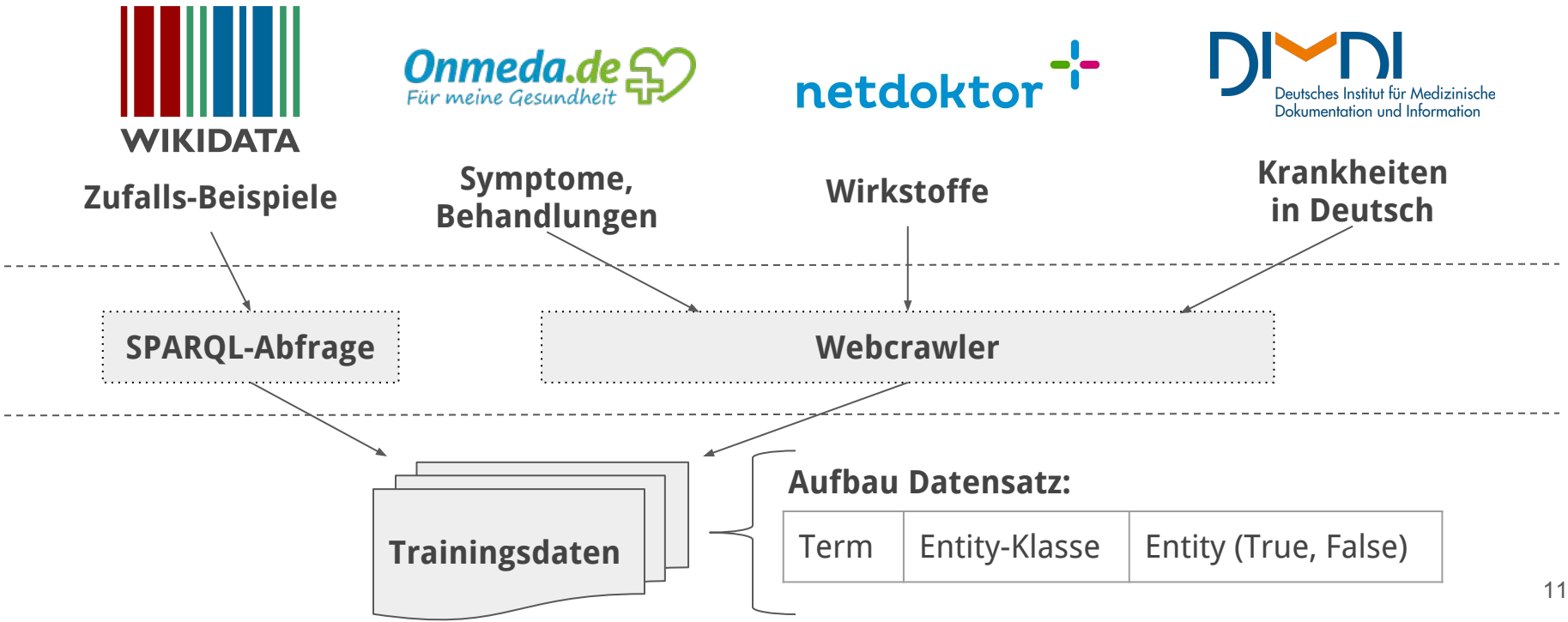
> Methodik

Die Grundlage des Encoders *bilden die Trainingsdaten*, mithilfe diesen kann der Encoder trainiert und verwendet werden



> Trainingsdaten beschaffen

Die Trainingsdaten wurden aus *verschiedenen Quellen* zusammengestellt



> Trainingsdaten beschaffen

Die *Qualität des Encoders* hängt stark von der *Qualität der gewählten Trainingsdaten* ab

	Term	**Entitäts-Klasse**	**Entität**
Beispiel Entity	Sonstiger Parasitenbefall der Haut	Krankheit	True
Beispiel nicht Entity	Zeitschrift für Politische Psychologie und Sexualökonomie	Zeitschrift	False

> Implementation NER Encoder

Der NER Encoder soll *3 verschiedene Merkmale* unterscheiden können

- Ist das Wort *eine medizinische Entität bzw. Teil einer* solchen Entität

  Parasitenbefall => **TRUE**
  Sexualökonomie => **FALSE**

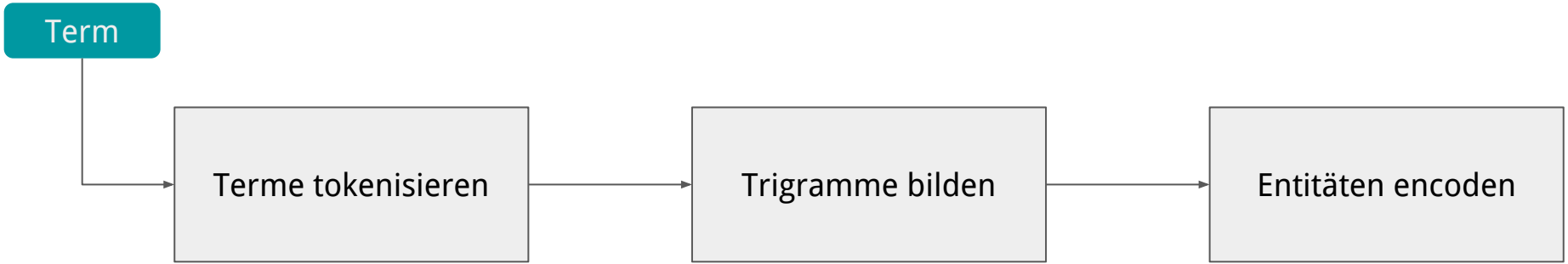
- In einem Satz mit einer medizinischen Entität, *wo im Satz*, steht das Wort? => BIOES-Tags

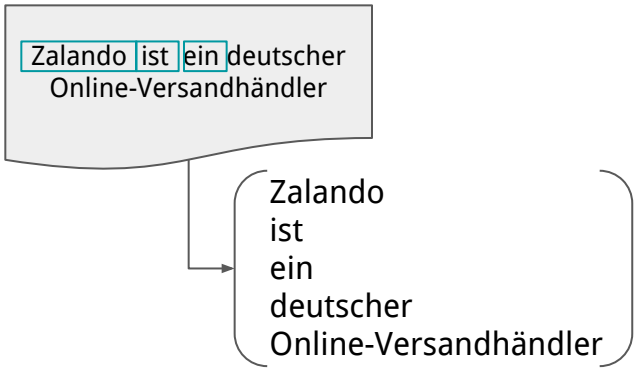
- Welcher *Entitäts-Klasse* gehört das Wort an?

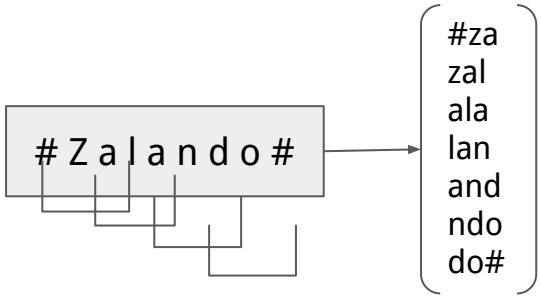
  Lungenkarzinom => **Disease**
  Chemotherapie => **Treatment**

> Implementation NER Encoder

Bei der Vorverarbeitung werden die Terme *in Tokens zerlegt* und *“trigramisiert”*







Labels:

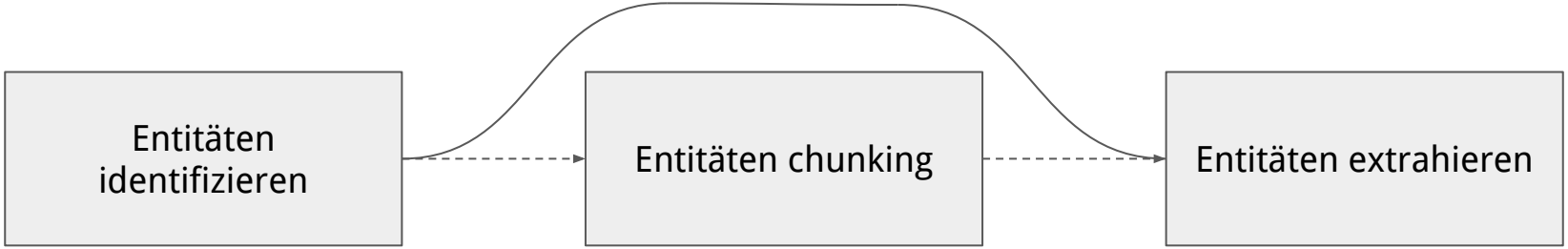
Entitäts-Klasse: Firma

BIOES-Tag: B

Entität? True

> Implementation NER Encoder

Nach der Vorverarbeitung werden *die Terme encoded* und *Entitäten extrahiert*. Mithilfe von den Tags können auch *Wordkombinationen* erkannt werden.



The most recent model is called the **Yeezy Boost 350 v2**

Entitäten

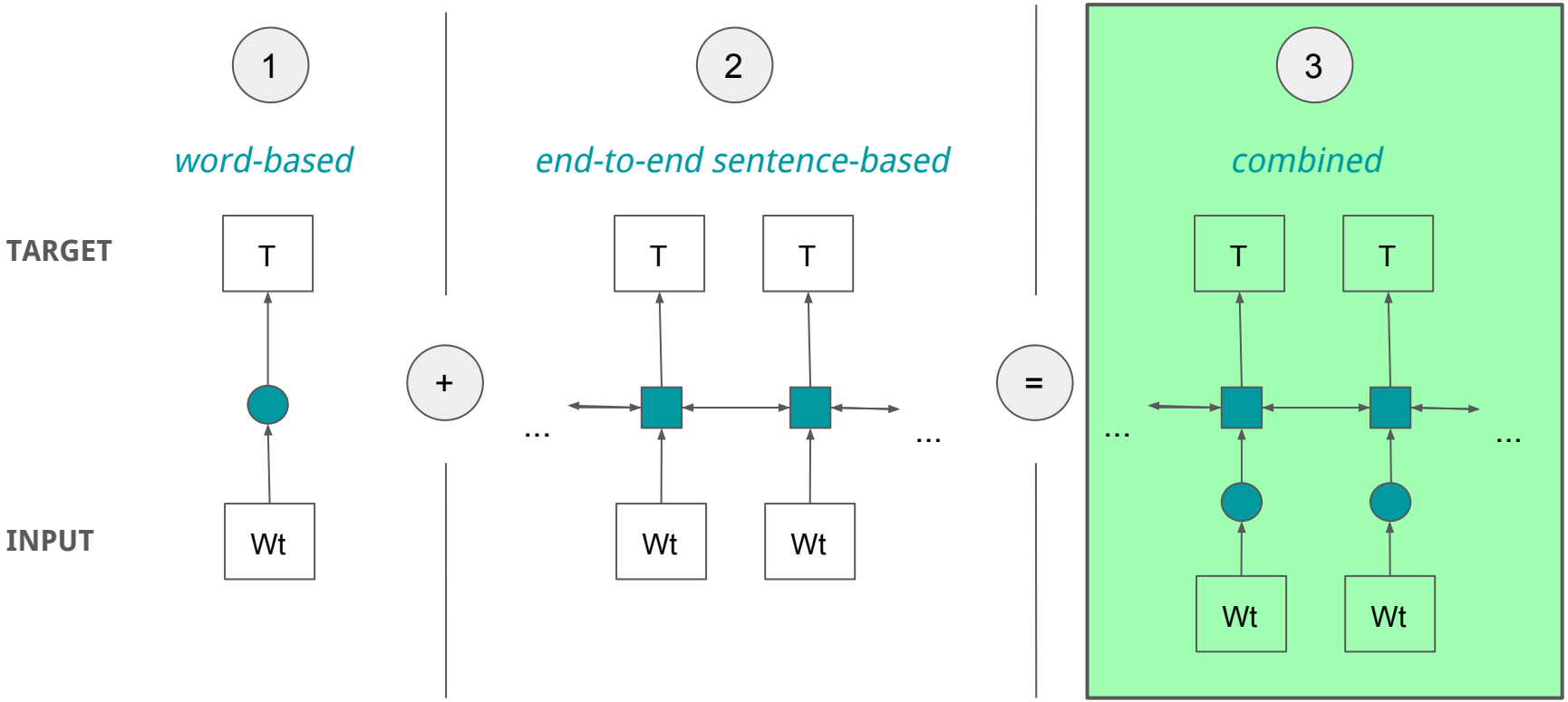
“Yeezy” + **“Boost”** + **“350”**+ **“v2”**

=> **Yeezy Boost 350 v2**

“Yeezy”, **“Boost”**, **“350”**, **“v2”**, **“Yeezy Boost 350 v2”**

=> **Schuh**

> Integration NER Encoder in Tasty



> Problemstellung

> Umsetzung

> Evaluation

> Fazit und Ausblick

> Evaluation

Die Evaluation unterteilt sich in 2 Teile:

1. Evaluation der *trainierten medizinischen Modelle*:
    - True/False Bestimmung der Entitäten
    - BIOES-Tagging
    - Klassifizierung der Entitäten in Entitätsklassen

2. Evaluation der *Integration in Tasty*:
    - BIOES-Tagging
    - Annotation von Entitäten

> Evaluation: Vergleich BIOES-Tagging von Tasty

Wir haben untersucht wie gut der trainierte *Entity Encoder* die verschiedenen *Tags disambiguieren* kann.

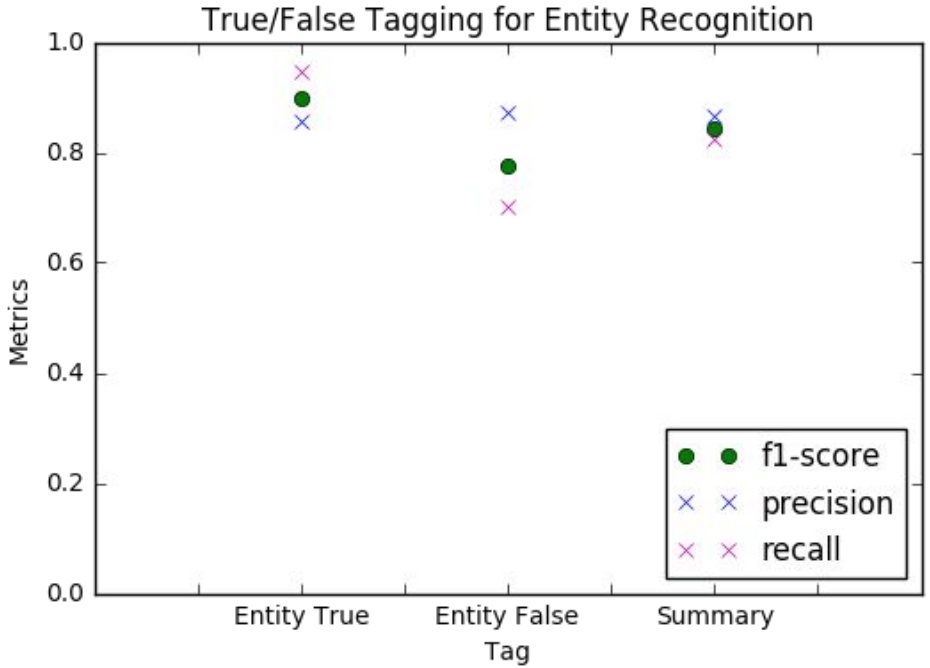
- Für das Training nutzten wir unsere selbst erstellten Datensätze:
  - 2005 Krankheiten
  - 228 Symptome
  - 264 Wirkstoffe
  - 242 Behandlungen
  - 2283 Random-Werte (Negativ-Beispiele)
    => 16’000 Tokens

- Im Test haben wir die Tokens mit der K-Fold Cross-Validation getestet
  - K wurde auf 5 gesetzt (Unterteilung in 5 Datensets)

> Evaluation: Analyse der True/False Tags

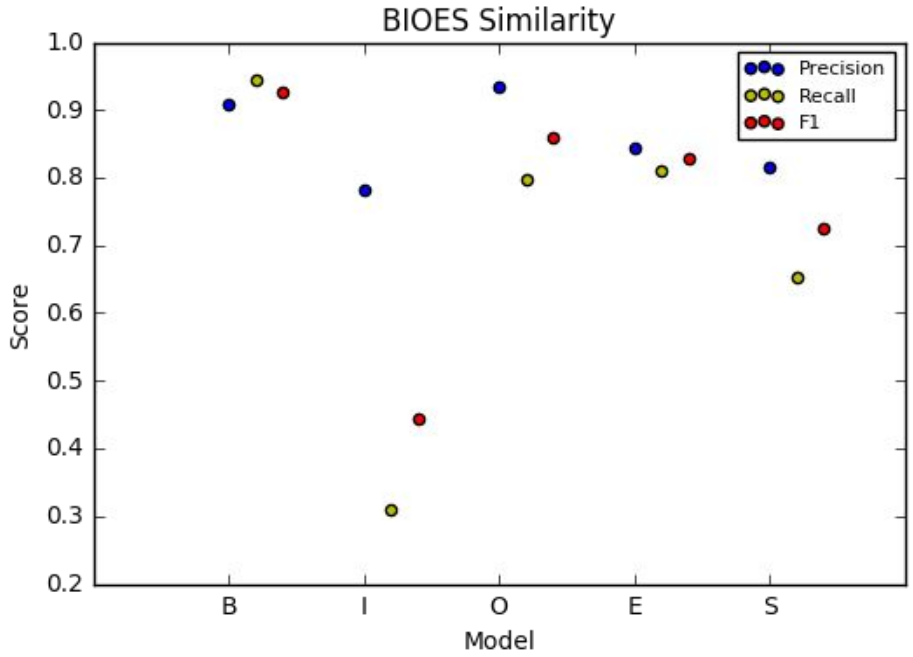
Beim True/False Tagging wurden die *Entitäten (Entity True) besser erkannt* als nicht Entitäten

=> Insgesamt wurde ein *F1-Score von 84%* erreicht



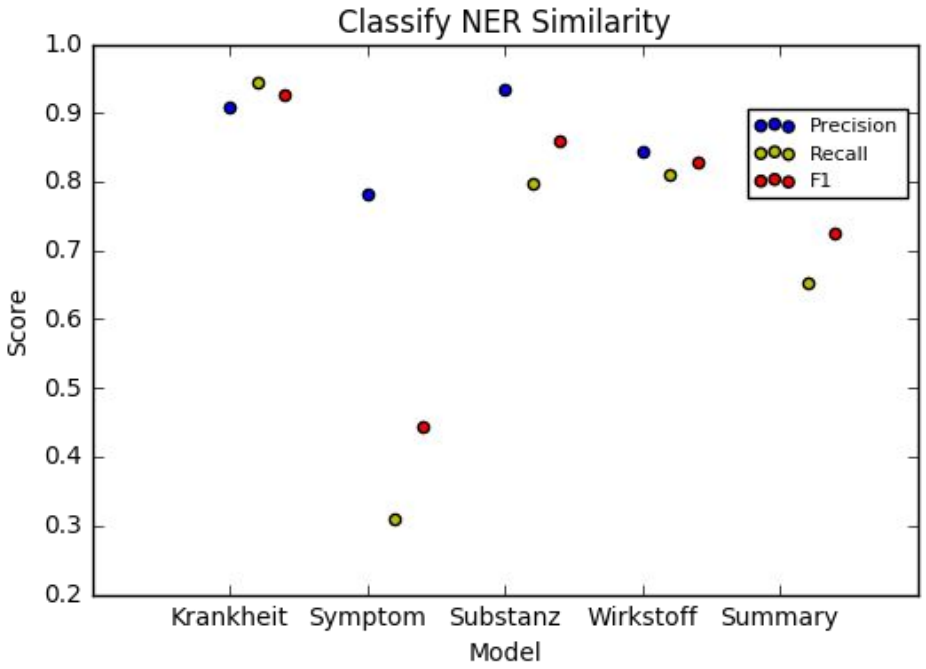
> Evaluation: Analyse der BIOES Tags

Beim BIOES-Tagging hatte der Encoder mit den *(I)nneren Elemente* der Entitäten am meisten Mühe, da viele Elemente *nicht erkannt wurden (Recall)*



> Evaluation: Analyse der Annotation der Entitätsklassen

Beim Entitätsklassen annotieren hatte der Encoder mit den *Symptomen* am meisten Mühe. Symptome kommen oft *auch in anderen Entitätsklassen vor*.



> Evaluation: Vergleich BIOES-Tagging von Tasty

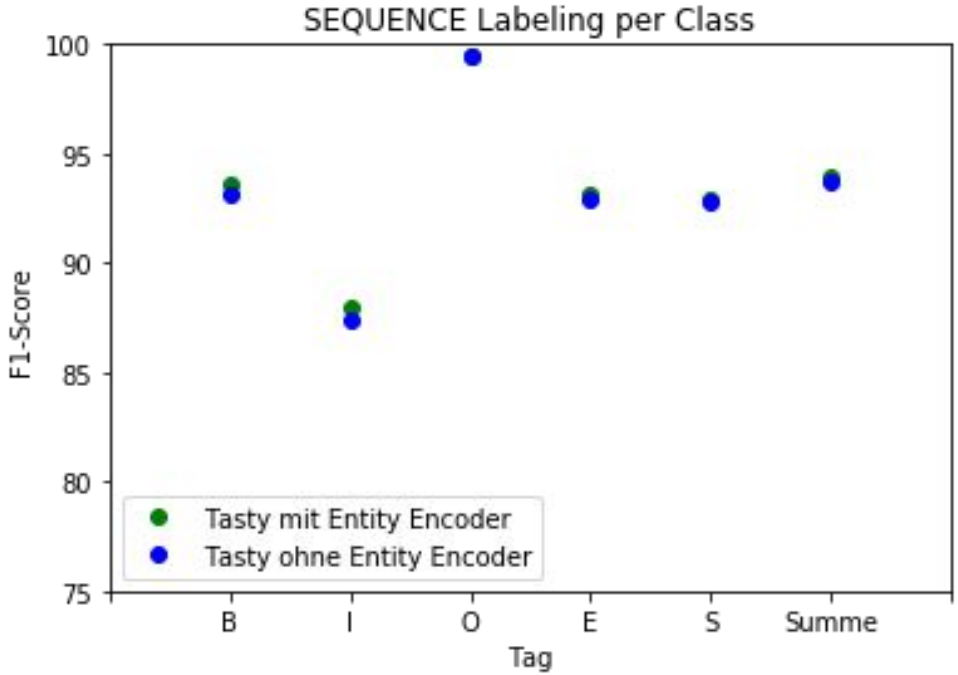
Wir haben untersucht ob wir *mit Hilfe von dem neuen Entity Encoder* für *True/False-Erkennung*, die Qualität von Tasty steigern können.

- Für das Training nutzten wir Wikipedia-Texte:
  - 5,291 Dokumente
  - 139,707 Sätze
  - 3,432,093 Tokens

- Im Test wurde versucht Titel aus Wikipedia in den Texten mit BIOES-Tags zu labeln. Der Test untersuchte:
  - 100 Dokumente
  - 2,754 Sätze
  - 67,513 Tokens

> Evaluation: Vergleich BIOES-Tagging von Tasty

Beim BIOES-Tagging wurden *Startelemente (B)* und *innere Elemente (I)* von Entitäten besser klassifiziert mit Hilfe des Entity Encoders



> Evaluation: Vergleich BIOES-Tagging von Tasty

Berechnet man *die Differenz der F1-Scores*, sieht man dies nochmals deutlicher

Alle Angaben in Prozenten:

B	I	O	E	S	Summe
+ **0.3925**	+ **0.5725**	- 0.005	+ 0.26	+ 0.03	+ 0.2725

> Evaluation: Vergleich Annotation von Tasty

In diesem Test wurde versucht *Titel aus Wikipedia als Entitäten* in den Texten *zu annotieren*.

- Der Test untersuchte:
  - 100 Dokumente
  - 67,513 Tokens
  - 5743 Entitäten sollten annotiert werden

> Evaluation: Vergleich Annotation von Tasty

Bei der Annotation wurden *mehr Entitäten erkannt und richtig annotiert*
=> höhere Accuracy
Jedoch auch ein paar *mehr Entitäten falsch annotiert* und
die Kombinationen der annotierten Tags war öfters falsch
=> tieferer F1-Score

Angaben in realen Werten:	**Ergebnis von Tasty mit Entity Encoder**	**Vergleich zu Tasty ohne Entity Encoder**
Gesamtanzahl annotierter Entitäten	**5738**	+ **47.0**
Anzahl *richtig annotierter* Entitäten	**5236** von 5743	+ **50.5**
Anzahl *falsch annotierter* Entitäten	**523**	+ 18.0
True-Accuracy	**98.64**	+ **0.025**
F1-Score	**90.915**	- 0.205

> Evaluation: Analyse schwer klassifizierbarer Spezialfälle

	True/False	Tag
Artikel/Bindewörter/Stop-Wörter		
die	**true**	**Disease**
der	**false**	**NotNER**
und	**false**	**NotNER**
Satzzeichen		
.	**true**	**Disease**
,	**false**	**NotNER**
(	**true**	**Disease**
)	**true**	**Symptome**

> Evaluation: Probleme durch Wortaufbau / Verwendung

	True/False	Tag
Fehler		
Asbest	**true**	**Disease**
Frau	**true**	**Symptome**

> Problemstellung

> Umsetzung

> Evaluation

> Fazit und Ausblick

> Fazit

Das Trainieren ist kein Problem mehr. Die *Überprüfung in großen Text-Korpora* jedoch schon.

- Für die Evaluation konnte das Problem mithilfe des WikiNER von Tasty umgangen werden, für **medizinische Texte fehlt** jedoch der notwendige **Gold-Standard**

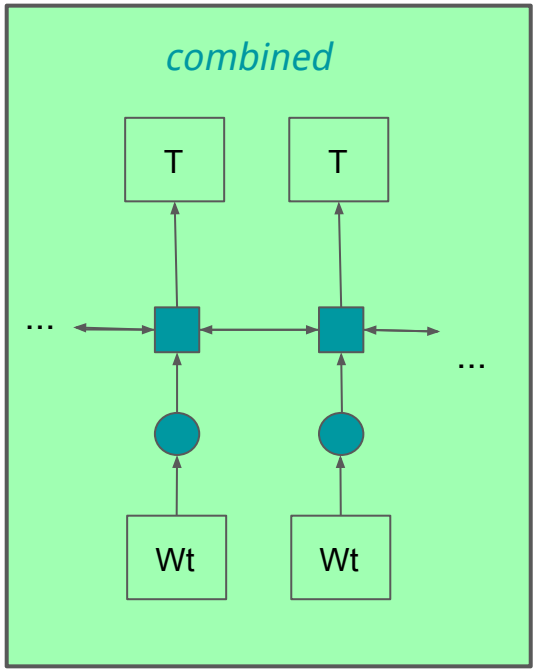
- **BIOES** Tags müssen durch den **Kontext** verstanden werden, ansonsten ist die Disambiguierung schwer

- Evaluation in Tasty hat **Potential aufgezeigt**
  - Tasty **funktioniert bereits sehr gut** auf getesteten Daten
  - Interessant wäre die **Analyse mit einem medizinischen, deutschen Text**

> Ausblick

Der Ansatz *sollte weiter verfolgt werden*. Der Encoder bietet gute Signale um Entities besser zu erkennen.

- Die Integration unseres Entity Encoders in TASTY
  - False-Positive Rate verbessern
- Gegen mehrere Wortlisten trainieren
- Testen gegen medizinischen, deutschen Gold-Standard
  - Gold-Standard manuell taggen



Gibt es noch Fragen zu Named Entity Recognition auf Basis von Wortlisten?