עיבוד שפה טבעית ש5:
תיוג רצפים - הבעיה

פרקים: SLP 8-8.3, Eisenstein 8

פתרון הבוחן

הגדרת הבעיה

- סיווג מסמכים
  - קלט: מסמך d שהוא רצף מילים
  - פלט: תג בודד y (יכול להיות קטגוריאלי או רציף)

- תיוג רצפים
  - קלט: מסמך d בעל m מילים
  - פלט: **m** תגים, כל אחד מהם בעל ערך מתוך קבוצה משותפת (או לעיתים, שונה)
    - השלכות על **טוקניזציה**

- דוגמאות
  - תיוג חלקי דיבר (כל מילה מקבלת תג) part-of-speech tagging
  - תיוג תכוניות מורפותחביריות (כל מילה מקבלת תגים והקלאסים נקבעים לפי חלק הדיבר שלה)
  - זיהוי ישויות (כל מילה מסומנת כחלק מישות או לא) named-entity recognition
  - (מידול שפה?????)

חלקי דיבר (Parts of Speech - POS)

- פועל Verb שם עצם Noun שם תואר Adjective מילת יחס Preposition …
- מגדירים **תפקידים אפשריים של מילה במשפט**
- מגדירים יחסי שכנות וקשר
  - תואר הפועל Adverb (*במהירות*, *בחוזקה*) - מתייחס לפעלים
  - באנגלית, תווית יידוע Determiner (למשל *a, the, some*) - חלה לפני שמות עצם (ותלוייהם)
- פתרון עמימות לגביהם הוא הרבה פעמים קריטי להבנה
  - עברית: **חולצה מטיילת** בוואדי; הדייג **דג** תפס
  - אנגלית: Squad helps dog **bite** victim, I saw **that** house music is popular

חלקי דיבר: פתוחים / סגורים (open-class / closed-class)

- חלק דיבר **סגור**: מילות שימוש, יש להן סמנטיקה אבל הן בעיקר "מסמנות" את מבנה המשפט (function words)
  - אנגלית: **the** parents **and their** children went **to the** zoo.
  - השוו עם עברית: **ה**הורים וילדי**הם** הלכו **ל**גן **ה**חיות
  - למה "סגור"? כי נדיר שמצטרפות מילים לקטגוריות האלו

- חלק דיבר **פתוח**: מילות "תוכן", קטגוריות שמתארות דברים בעולם ומילים חדשות יכולות להצטרף כשהעולם משתנה (content words)
  - פרויקטור ה**קורונה קימפל** כללי **ריצוף** חדשים

Open class ("content") words

Nouns

Proper
Janet
Italy

Common
cat, cats
mango

Verbs

Main
eat
went

Auxiliary
can
had

Adjectives *old green tasty*

Adverbs *slowly yesterday*

Numbers
122,312
one

Interjections *Ow hello*

… more

Closed class ("function")

Determiners *the some*

Conjunctions *and or*

Pronouns *they its*

Prepositions *to with*

Particles *off up*

… more

(מתוך השקפים של סטנפורד, קרדיט דן ג'ורפסקי)

עד כמה אוניברסליים חלקי הדיבר?

- הקורפוס המסיבי הראשון באנגלית, Penn Treebank (PTB), כולל 36 חלקי דיבר ברזולוציה גבוהה
  - למשל, VBG - Verb, gerund or present participle
  - או RBS - Adverb, superlative
  - או WP$ - Possessive wh-pronoun
  - שלושתם לא קיימים בעברית ברמת המילה

- לפעמים אפילו ה"בסיסיים" לא תמיד רלוונטיים
  - ברוסית אין תוויות יידוע
  - בעברית אין particles (כמו up בביטוי wake up)
  - בקוריאנית מילות יחס באות אחרי המשלים (ולכן אינן **pre**positions)

חלקי דיבר "אוניברסליים"

- כבר PTB "מהונדס" כך שניתן לחלץ תגים בלתי-מעודנים (coarse-grained) מהאות או שתיים הראשונות בשם התג העדין (V - verb, N - noun וכו')

- ב-2012 יצא סט אוניברסלי-בשאיפה עם 17 תגים
  - לרבות תג "תופס-כל" X
  - בשימוש גם לתופעות כמו עירוב שפות

- ועדיין למטרות ספציפיות מפתחים סטי-תגים מיוחדים
  - למשל, טוויטר

N	common noun
O	pronoun (personal/WH; not possessive)
^	proper noun
S	nominal + possessive
Z	proper noun + possessive
V	verb including copula, auxiliaries
L	nominal + verbal (e.g. *i'm*), verbal + nominal (*let's*)
M	proper noun + verbal
A	adjective
R	adverb
!	interjection

ikr	smh	he	asked	fir	yo	last
!	G	O	V	P	D	A
name	so	he	can	add	u	on
N	P	O	V	V	O	P
fb	lololol					
^	!					

E	emoticon
$	numeral
,	punctuation
G	other abbreviations, foreign words, possessive endings, symbols, garbage

	Tag
Open Class	ADJ
	ADV
	NOUN
	VERB
	PROPN
	INTJ
Closed Class Words	ADP
	AUX
	CCONJ
	DET
	NUM
	PART
	PRON
	SCONJ
Other	PUNCT
	SYM
	X

חלקי דיבר: גזירה מול הטיה (derivation vs. inflection)

- עברית: עמד / עמדה / עומדים / עומדות
  - אותו חלק דיבר, אותו ערך מילוני
  - ההבדלים בין הצורות הן של **התאם**
- קם / מקום / מיקם
  - חלקי דיבר שונים, ערך מילוני נפרד (= lemma נפרדת)
  - סמנטיקה גמישה, שימושים מטאפוריים אפשריים
- אנגלית: mark / mark, או bite / bite
  - "גזירת אפס" zero derivation

earnings growth took a back/ADJ seat
a small building in the back/NOUN
a clear majority of senators back/VERB the bill
enable the country to buy back/PART debt
I was twenty-one back/ADV then

- גזירה מיוצגת ע"י תיוג חלקי דיבר, מה לגבי הטיה?

תיוג מורפותחבירי (morphosyntactic attribute tagging)

- סימון נוסף ברמת המילה עבור תכוניות ההתאם
- שונה משפה לשפה, ומתוחזק עבור כל חלק דיבר בנפרד

```
# text = תופעה זו התבררה אתמול בוועדת העבודה והרווחה של הכנסת, שדנה בנושא העסקת עובדים זרים.
1	תופעה	תופעה	NOUN	NOUN	Gender=Fem|Number=Sing
2	זו	זו	PRON	PRON	Gender=Fem|Number=Sing|Person=3|PronType=Dem
3	התבררה	התברר	VERB	VERB	Gender=Fem|HebBinyan=HITPAEL|Number=Sing|Person=3|Tense=Past
4	אתמול	אתמול	ADV	ADV	_
```

תיוג רצפים - מטריקות (I)

- ניחוש ראשון: דיוק ברמת המילה (accuracy)
- בפועל, כמובן, אין דין כל חלק דיבר כמשנהו
  - (מה הדיוק של most common tag baseline על PTB, מה התוצאה הכי טובה?)
- יש שמדווחים דיוק עבור חלקי דיבר פתוחים בלבד
- או שמאחדים כמה חלקי דיבר יחד
  - מה קורה אם מאחדים במטריקה אבל לא באימון, לעומת גם וגם?
- עוד מטריקה אפשרית - F1 על כל התגים

- מה לגבי תגים מורפותחביריים?
  - (מיקרו מול מאקרו)

עברית קשה שפה

.עשרות אנשים מגיעים מתאילנד לישראל כשהם נרשמים כמתנדבים, אך למעשה משמשים עובדים שכירים זולים

1	עשרות	NUM
2	אנשים	NOUN
3	מגיעים	VERB
4–5	מתאילנד	_
4	מ	ADP
5	תאילנד	PROPN
6–7	לישראל	_
6	ל	ADP
7	ישראל	PROPN
8–9	כשהם	_
8	כש	SCONJ
9	הם	PRON
10	נרשמים	VERB
11–12	כמתנדבים	_
11	כ	ADP
12	מתנדבים	NOUN
13	,	PUNCT
14	אך	CCONJ
15	למעשה	ADV
16	משמשים	VERB
17	עובדים	NOUN
18	שכירים	ADJ
19	זולים	ADJ
20	.	PUNCT

איסור על מתן שכר לעובדים מתחת לשכר המינימום

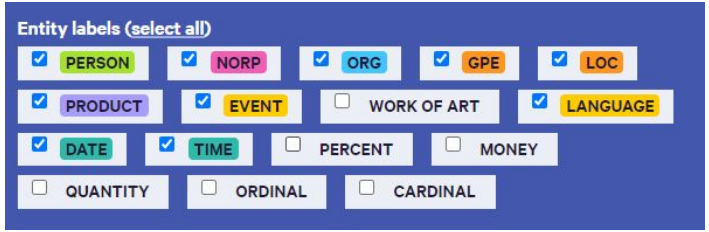
22–24	לעובדים
22	ל
23	ה_
24	עובדים

זיהוי ישויות NER

David Ben-Gurion PERSON (16 October 1886 DATE – 1 December 1973 DATE) was the primary national founder of the State of Israel ORG and the first prime minister of Israel GPE . Adopting the name of Ben-Gurion ORG in 1909 DATE , he rose to become the preeminent leader of the Jewish NORP community in British NORP -ruled Mandatory Palestine GPE from 1935 DATE until the establishment of the State of Israel ORG in 1948 DATE , which he led until 1963 DATE with a short break in 1954–55.

He stepped down from office in 1963 DATE , and retired from political life in 1970 DATE . He then moved to Sde Boker PERSON , a kibbutz in the Negev LOC desert, where he lived until his death. Posthumously, Ben-Gurion PERSON was named one of Time ORG magazine's 100 Most Important People of the 20th century DATE .

SpaCy



זיהוי ישויות - שימושים

- מענה על שאלות (Question Answering): התשובה היא ישות
- חילוץ מידע (Information Extraction): מציאת מידע על העולם שניתן לעדכן בעזרתו מסד נתונים מסודר
- ניתוח סנטימנט יחסי: מה דעתה של הכותבת לגבי חברה / מוצר / אדם כלשהו

זיהוי ישויות - תיוג

- ניחוש ראשון - בינארי (האם מילה היא חלק מישות)
- איך נתייג ברמת המילה:
  את כל כתבי וואלה ערן טיפנבורן מנהל.
- פתרון אחד (שגם רלוונטי להגדרת הבעיה) - סוגי ישויות ORG, PER, LOC
  - אוקיי, אבל מה עם: *את כל כתבי וואלה ידיעות אחרונות שכרו הבוקר* ?
- פתרון מבני יותר - סכמת תיוג (BIO)
  - תחילת ישות B, אמצע ישות I, לא חלק מישות O

I	PRON	O	
live	VERB	O	
in	ADP	O	
United	PROPN	B	GPE
states	PROPN	I	GPE
of	ADP	I	GPE
America	PROPN	I	GPE
,	PUNCT	O	
my	DET	O	
Phone	NOUN	O	
number	NOUN	O	
is	AUX	O	
(123)123-1234	CD	B	

זיהוי ישויות - תיוג

- ניחוש ראשון - בינארי (האם מילה היא חלק מישות)
- איך נתייג ברמת המילה:
  את כל כתבי וואלה ערן טיפנבורן מנהל.
- פתרון אחד (שגם רלוונטי להגדרת הבעיה) - סוגי ישויות ORG, PER, LOC
  - אוקיי, אבל מה עם: *את כל כתבי וואלה ידיעות אחרונות שכרו הבוקר* ?
- פתרון מבני יותר - סכמת תיוג (BIO)
  - תחילת ישות B, אמצע ישות I, לא חלק מישות O
  - שיטה אחרת: BILOU / BIOES

I	PRON	O	
live	VERB	O	
in	ADP	O	
United	PROPN	B	GPE
states	PROPN	I	GPE
of	ADP	I	GPE
America	PROPN	I	GPE
,	PUNCT	O	
my	DET	O	
Phone	NOUN	O	
number	NOUN	O	
is	AUX	O	
(123)123-1234	CD	B	

זיהוי ישויות - תיוג

- האם פתרנו את כל בעיות התיוג?
  - מה לגבי: *The New York Times Advisory Board*? *הוועד הפועל של רשות השידור*?

זיהוי ישויות - מטריקות

- תרגמנו את הבעיה לתיוג רצפים, אבל האם נרצה למדוד ביצועים ברמת המילה?

- מדידה **ברמת היישות**
  - טעינו איפשהו בטווח המילים של הישות (span) ⇐ טעות בכל הישות
  - מקל על חישוב Recall ו-Precision
  - (מה החשיבות של להצליח בחיזוי מילה שהיא O?)
  - ציון F1 פר כל הישויות / פר סוג ישות
    - מיקרו? מאקרו?

- התפלגות סוגי ישויות בדאטאסט של ציוצים (WNUT 2016):

	Train	Dev	Test
company	171	39	621
facility	104	38	253
geo-loc	276	116	882
movie	34	15	34
musicartist	55	41	191
other	225	132	584
person	449	171	482
product	97	37	246
sportsteam	51	70	147
tvshow	34	2	33
Total	1496	661	3473

תיוג רצפים - פירמול מדויק יותר

- קלט: מסמך d בעל m מילים
- פלט: m תגים, כל אחד מהם בעל ערך מתוך קבוצה משותפת (או לעיתים, שונה)

- אבל התגים תלויים אחד בשני, גם בחלקי דיבר אבל **בוודאי** שבזיהוי ישויות
- אנחנו בעולם של **חיזוי מבני Structured Prediction**

- **מה סט הקלאסים האמיתי של כל המסמך (Y)?**

- בצורתו החופשית ביותר, ה-Y הפוטנציאלי הוא אוסף השמות התגים הנכונים עבור **כל m המילים**. כלומר, Y(m).

דוגמא

- אוסף כל התגים האפשריים - N,V,A,P

- x = גנן גידל דגן
- Y(3) = {<N,N,N>, <N,N,V>, <N,N,A>, <N,N,P>, **<N,V,N>**, <N,V,V>, …, …, <P,A,A>, <P,A,P>, <P,P,N>, <P,P,V>, <P,P,A>, <P,P,P>}

תיוג רצפים - פירמול

- קלט: מסמך d בעל m מילים
- פלט: m תגים, כל אחד מהם בעל ערך מתוך קבוצה משותפת (או לעיתים, שונה)

- אבל התגים תלויים אחד בשני, גם בחלקי דיבר אבל **בוודאי** שבזיהוי ישויות
- אנחנו בעולם של **חיזוי מבני Structured Prediction**

- **מה סט הקלאסים האמיתי של כל המסמך (Y)?**

- בצורתו החופשית ביותר, ה-Y הפוטנציאלי הוא אוסף השמות התגים הנכונים עבור **כל m המילים**. כלומר, Y(m).
- לכן, שיטות לפתרון הבעיה יתמודדו לא רק עם הסקה (inference) ועם למידה (learning), אלא גם עם **חיפוש** (search).

למחר - יהיה נחמד אם תדפיסו את המטריצות מהמודל

- (או שיהיו נגישות על מסך)

Transitions	-> V	-> N	-> M	-> P	-> J
V ->	0.05	0.35	0.25	0.2	0.15
N ->	0.6	0.05	0.05	0.1	0.2
M ->	0.6	0.1	0.1	0.15	0.05
P ->	0.05	0.5	0.3	0.01	0.14
J ->	0.5	0.05	0.1	0.2	0.15
START ->	0.15	0.25	0.25	0.3	0.05

Emissions	שרה	שיר	שמח
V	0.02	0.002	0.001
N	0.005	0.01	0
M	0.03	0.02	0.001
P	0	0.00001	0
J	0	0	0.04

Transitions (log)	-> V	-> N	-> M	-> P	-> J
V ->	-2.996	-1.050	-1.386	-1.609	-1.897
N ->	-0.511	-2.996	-2.996	-2.303	-1.609
M ->	-0.511	-2.303	-2.303	-1.897	-2.996
P ->	-2.996	-0.693	-1.204	-4.605	-1.966
J ->	-0.693	-2.996	-2.303	-1.609	-1.897
	-1.897	-1.386	-1.386	-1.204	-2.996

Emissions (log)	שרה	שיר	שמח
V	-3.912	-6.215	-6.908
N	-5.298	-4.605	
M	-3.507	-3.912	-6.908
P		-11.513	
J			-3.219