A New Approach for Named Entity Recognition

Burak Ertopçu[1], Ali Buğra Kanburoğlu[1], Ozan Topsakal[1], Onur Açıkgöz[1], Ali Tunca Gürkan[1], Berke Özenç[1], İlker Çam[1], Begüm Avar[2], Gökhan Ercan[1], Olcay Taner Yıldız[1]

[1]Department of Computer Engineering, Işık University, İstanbul, Turkey
[2]Department of Linguistics, Boğaziçi University, İstanbul, Turkey

August 25, 2017

Outline

Definition

Anything that is denoted by a proper name, i. e., for instance, a person, a location, or an organization, is considered to be a named entity.

[$_{ORG}$ Türk Hava Yolları] bu [$_{TIME}$ Pazartesi'den] itibaren [$_{LOC}$ İstanbul] [$_{LOC}$ Ankara] güzergahı için indirimli satışlarını [$_{MONEY}$ 90 TL'den] başlatacağını açıkladı.

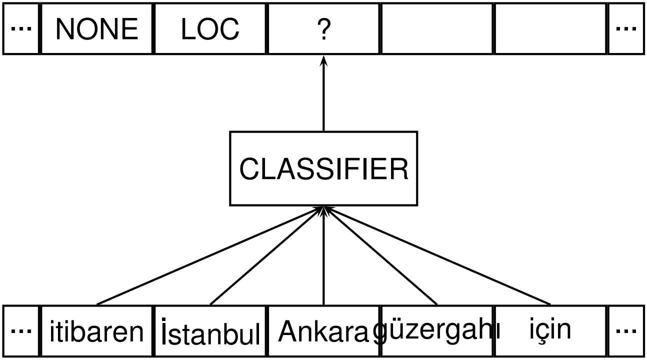
Ner Categories

Tag	Sample Categories	Example
PER	people, characters	**Atatürk** yurdu düşmanlardan kurtardı.
ORG	companies, teams	**IMKB** günü 60 puan yükselerek kapattı.
LOC	regions, mountains, seas	Ülkemizin başkenti **Ankara'dır**.
TIME	time expressions	**Cuma günü** tatil yapacağım.
MONEY	monetarial expressions	Geçen gün **3000 TL** kazandık.

Named entity tagging as classification problem

Word	Features				Label
	Root	Pos	Capital	...	
Türk	Türk	Noun	True	...	ORGANIZATION
Hava	Hava	Noun	True	...	ORGANIZATION
Yolları	Yol	Noun	True	...	ORGANIZATION
bu	bu	Pronoun	False	...	NONE
Pazartesi'den	Pazartesi	Noun	True	...	TIME
itibaren	itibaren	Adverb	False	...	NONE
İstanbul	İstanbul	Noun	True	...	LOCATION
Ankara	Ankara	Noun	True	...	LOCATION
güzergahı	güzergah	Noun	False	...	NONE
için	için	Adverb	False	...	NONE
indirimli	indirimli	Adjective	False	...	NONE
satışlarını	sat	Noun	False	...	NONE
90	90	Number	False	...	MONEY
TL'den	TL	Noun	True	...	MONEY
başlatacağını	başlat	Noun	False	...	NONE
açıkladı	açıkla	Verb	False	...	NONE
.	.	Punctuation	False	...	NONE

Named entity tagging as classification problem



itibaren	İstanbul	Ankara	güzergahı	için
Adverb	Noun	Noun	Noun	Adverb
False	True	True	False	False

Word Embeddings

- Traditional representations of words (i.e., one-hot vectors) are based on word-word ($W \times W$) co-occurrence sparse matrices.
- Distributed word representations (DRs) (i.e., word embeddings) are word-context ($W \times C$) dense matrices.
- DRs are real valued vectors where each context can be considered as a continuous feature of a word.

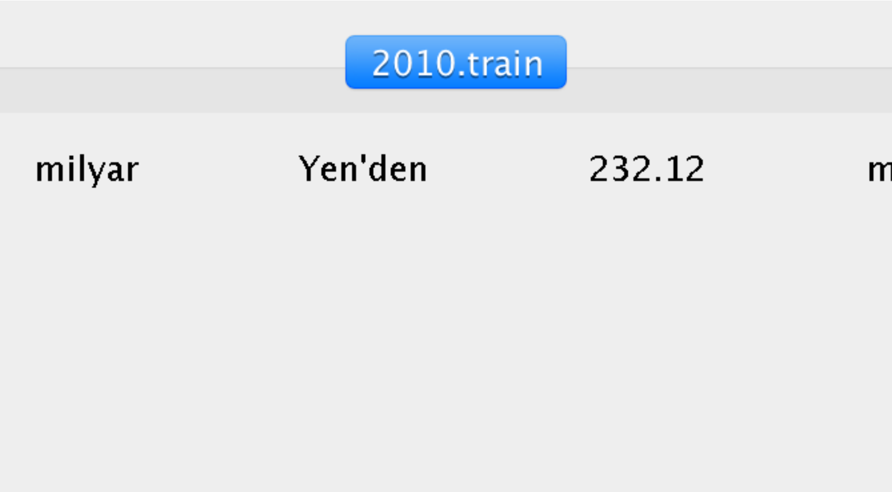
Mikolov's Work

- Mikolov et al.'s SkipGram is an unsupervised neural network based distributional semantic model (DSM).
- Main idea is to learn distributed word representations through maximizing the probability of surrounding words within a window by learning weights of each word vector in context dimensions.
- Continuous bag-of-word (CBOW) model is also proposed by Mikolov et al. which is reported to be more scalable than SkipGram model.
- While SkipGram predicts surrounding words of a current word w, CBOW model predicts the w based on the context.

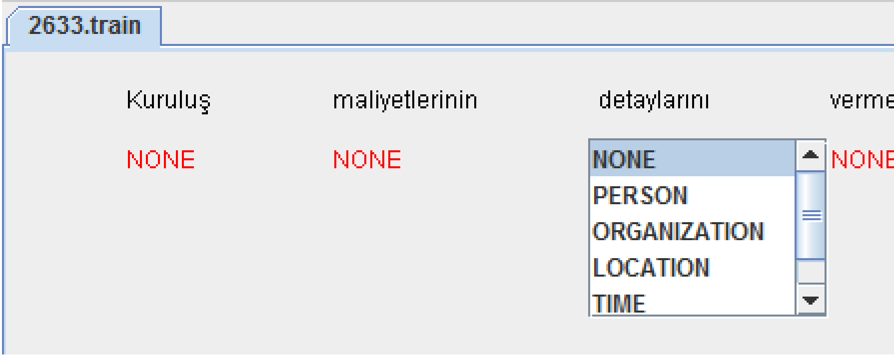
- Collected from Penn-Treebank corpus and each sentence of this dataset is translated into Turkish.
- 1400 sentences, 13194 words.

Label	Count
PERSON	606
LOCATION	235
ORGANIZATION	685
MONEY	387
TIME	299
NONE	10982

Morphological Disambiguation



Ner Tagging


2633.train
Kuruluş
maliyetlerinin
detaylarını
NONE
NONE
NONE
PERSON
ORGANIZATION
LOCATION
TIME
NONE

Classification Algorithms

- Dummy: Decides based on the prior class probability without looking at the input.
- C45: The archetypal decision tree method.
- Knn: K-Nearest Neighbor classification algorithm.
- Lp: Linear perceptron.
- Mlp: Well-known multilayer perceptron.
- Nb: Classic Naive Bayes classifier.
- Rf: Random Forest method.

Discrete Model: Features

- CaseAttribute (`C`)
- IsCapitalAttribute (`IC`)
- IsDateAttribute (`ID`)
- IsFractionAttribute (`IF`)
- IsHonorificAttribute (`IH`)
- IsMoneyAttribute (`IM`)
- IsNumAttribute (`IN`)
- IsOrganizationAttribute (`IO`)
- IsPropAttribute (`IP`)
- IsRealAttribute (`IR`)
- IsTimeAttribute (`IT`)
- MainPosAttribute (`MP`)
- LastIGContainsPossessiveAttribute (`P`)
- RootFormAttribute (`RF`)
- RootPosAttribute (`RP`)
- SurfaceFormAttribute (`SF`)

Discrete Model: Methods

Method	Attributes	W	C	Parameters
METHOD1	IC, ID, IR, IT, MP	4	Rf	M = 4, N = 20
METHOD2	IC, ID, IF, IT, MP, RF, SF	1	Mlp	μ = 0.1, h = 30
METHOD3	IH, IM, MP, RF, SF	0	Mlp	μ = 0.1, h = 50
METHOD4	C, IC, IH, IO, IR, MP, P, RF, RP, SF	2	Lp	μ = 0.1
METHOD5	IC, IN, IP, MP, RF, SF	1	Mlp	μ = 0.1, h = 5, 8
METHOD6	IC, IP, MP, RF, SF	1	Mlp	μ = 0.1, h = 10

Continuous Model: Hyperparameters

- Context window (`win`): Word context window size where co-occurrence information is gathered. Default is 5.
- Dimension (`d`): Dimension size of the word embeddings. Number of neurons in neural network layers. Default is 100.
- Deleting infrequent words (`del`): Threshold frequency value for excluding words from the training. Default is 0.

Continuous Model: Word Forms

- Surface Form (SU): Natural form of a word which appeared in a text as it is. Ex: *Güzel gözlü turnalar, göçtüler.*
- Root Form (RO): Root of a word used in DR training based on morphological disambiguation of every sentence. Ex: *Güzel göz turna, göç.*

Inter-annotator Aggrement

- Two different group of annotators annotated same sentences.
- We could only re-annotate 100 of the total of 1400 sentences.
- %97.5 inter-annotator aggrement.
- Expected inter-annotator agreement is %16.7.

Discrete Model

Classifier	Error Rate
DUMMY	14.89
METHOD1	14.71
METHOD2	7.64
METHOD3	12.19
METHOD4	7.65
METHOD5	8.45
METHOD6	9.28

Continuous Model

	OOV%	Dummy	Lp	Mlp	Nb	Knn	C45
SURFACE (SU)							
100K	32.4	9.1	8.13	8.25	35.79	9.4	9.13
500K	23.05	10.75	8.06	7.81	35.18	10.24	10.61
1M	20.31	11.18	8.05	7.94	31.57	10.23	9.96
ROOT (RO)							
100K	20.77	9.79	6.87	6.74	15.22	9.3	7.85
500K	17.49	11.1	6.96	6.62	20.43	14.8	9.55
1M	16.33	11.56	6.7	6.56	9.52	16.82	9.87
OTHER CONF.							
1M-RO-d300	16.33	11.56	6.96	6.65	10.04	21.67	9.54
1M-RO-d20	16.33	11.56	7.47	6.76	11.49	7.92	9.33
1M-RO-d10	16.33	11.56	8.3	7.24	13.13	7.4	10.55
MIN	16.33	9.1	6.7	6.56	9.52	7.4	7.85

Questions?