Named Entity Labeling Manuel for Suder Turkish News Corpora

March 22, 2020

We propose a labeling strategy for the named entities encountered in Suder Turkish News Corpora. We have a label set for named entities

$$\{pb, pi, lb, li, ob, oi, tb, ti, mb, mi, r\}$$

with size 11. First letter of the tag stands for the type of the entity such that p for person, l for location, o for organization, t for time, m for money and r for other. Second letter of the tag represents the place of the token for the entity that contains it. b means "begining" of the entity, i means "in" the entity. First token of the entity will have a label with b, tokens after that will have a tag with i. For example `World Health Organization` will have labels ob, oi, oi respectively.

Named entities have to be labeled considering following rules;

1. Proper nouns that indicate a name of a person must be labeled as pb or pi, e.g. John: pb Doe: pi.
2. Appellations and titles are not included in the person entity, e.g. Prof.: r John: pb Doe: pi.
3. All addresses, country names, geographical names, continent names, city names and exact locations specified with a proper noun must be labeled as lb or li.
4. All institutions, organizations, companies, schools, hospitals, foundations, governmental organizations, sport clubs etc. must be labeled as ob or oi. If a proper noun indicates a structure which have employees, an organizational structure and a function must be labeled as ob or oi.
5. All time indicators with a formal format must be labeled as tb or ti. For example, tuesday: r, second: r friday: r of: r the: r March: r, twenty: r minute: r after: r the: r midnight: r, next: r April: r but 5:30pm: tb, 5.30am: tb , 12.12.2007: tb, December: tb 1999: ti, 1st: tb of: ti August: ti, 5: tb pm: ti.

6. All money indicators with a exact amount and currency must be labeled as *mb* or *mi*. For instance, lots: *r* of: *r* money: *r*, one: *mb* hundred: *mi* dollar: *mi*, 100: *mb* dollar: *mi*, 100: *mb* $: *mi*, A: *r* Fistful: *r* of: *r* Dollars: *r*.

Following explanations and examples will set the boundaries between named entities;

- Proper nouns can not be seperated into two different named entities, e.g. Vienna: *lb* International: *ob* Film: *oi* Festival: *oi* is wrong, it should be labeled as Vienna: *ob* International: *oi* Film: *oi* Festival: *oi*.
- Nation names and ethnic indicators are not named entities, if they are alone, e.g. Turk: *r* but Turk: *ob* Silahlı: *oi* Kuvvetleri: *oi*.
- Noun clauses should be evaluated seperately if they do not obey any of the mentioned rules, e.g. Mısır: *lb* Kralı: *r*, Rize: *lb* Belediye: *r* Başkanı: *r* Mustafa: *pb* Satıcı: *pi*.
- Adjective clauses must be labeled seperately if they are not part of a proper noun, e.g. Eski: *r* İstanbul: *lb* but İstanbul: *ob* Büyük: *oi* Şehir: *oi* Belediyesi: *oi*.
- Some other examples; Ankara: *ob* 9.: *oi* Ağır: *oi* Ceza: *oi* Mahkemesi: *oi*, 18: *lb* Mart: *li* Stadyumu: *li*, Suriyeli: *r*, Ahmet: *pb* T.: *pi*, TBMM: *ob*, MHP: *ob*, Radisson: *ob* Blu: *oi* Oteli: *oi*, Zeytinburnu: *lb* Kozlu: *lb* Mezarlığı'ndaki: *lb* gasilhaneden: *r*, Sabri: *ob* Ülker: *oi* Gıda: *oi* Araştırmaları: *oi* Enstitüsü: *oi* Vakfı: *oi*, milli: *r* takım: *r*, 15: *tb* temmuzdan: *ti* 2: *r* gün: *r* sonra: *r*.