Exercise: Entity linking

November 8, 2016

Throughout this excercise, you will annotate a sample text using simple (yet effective) entity linking approach, known as "CMNS"[1].
We provide you with the data from the knowledge graph and ask you to annotate the documents within the general pipeline of entity linking, that is mention detection, entity ranking, and disambiguation. Table 1 presents an excerpt from a surface form dictionary, together with the number of times an entity appeared as the target link of the mention in Wikipedia (denoted as *count*). *"_total"* is the total number of times a mention is linked to any entity.

Sample text:

"... Angola changed from a one-party Marxist-Leninist system ruled by the MPLA to a formal multiparty democracy following the 1992 elections ..."

Table 1: An excerpt from the surface form dictionary.

Mention	**Entity**	**Count**
1992 elections	<wikipedia:Philippine_general_election,_1992>	9
1992 elections	<wikipedia:Angolan_presidential_election,_1992>	1
1992 elections	*_total*	98
angola	<wikipedia:Angola>	4026
angola	<wikipedia:Angola_(Portugal)>	6
angola	<wikipedia:Angola_national_football_team>	120
angola	*_total*	4298
democracy	<wikipedia:Democracy>	108
democracy	<wikipedia:Democracy_(album)>	3
democracy	*_total*	2162
multiparty democracy	<wikipedia:multiparty_democracy>	11
multiparty democracy	*_total*	11
one party	<wikipedia:Non-possessors>	1
one party	<wikipedia:Single-party_state>	5
one party	*_total*	983

[1] pronounced as commonness.

Step 1: Mention detection

Mention detection in CMNS is based on the following heuristics:

It starts with longest possible n-gram of the text (e.g. $n = 8$). If the n-gram is found in the dictionary, the mention and the corresponding entities are kept (and the shorter n-grams are ignored). Otherwise, it tries to match the (n-1)-grams. The algorithm continues recursively until a mention is found or n reaches to 1.

Question. Considering Table 1, what is the output of the mention detection step for the given sample text?

Answer. All mention-entity pairs of Table 1 are considered, except the ones related to the mention “democracy“. We ignore this mention, because the longer mention “multiparty democracy“ is considered.

Step 2: Entity ranking

Entity ranking in CMNS is based on the commonness score:

$$Commonness(e, m) = p(e|m) = \frac{n(m, e)}{\sum_{e'} n(m, e')}, \tag{1}$$

where $n(m, e)$ denotes the number of times entity e is the link target of mention m.

Question. Compute the commonness for all mention-entity pairs, where mention is “1992 elections”.

Mention	**Entity**	**Commonness**
1992 elections	<wikipedia:Philippine_general_election,_1992>	$9/98 = 0.09$
1992 elections	<wikipedia:Angolan_presidential_election,_1992>	$1/98 = 0.01$
angola	<wikipedia:Angola>	$4026/4298 = 0.93$
angola	<wikipedia:Angola_national_football_team>	$120/4298 = 0.03$
angola	<wikipedia:Angola_(Portugal)>	$6/4298 = 0.001$
multiparty democracy	<wikipedia:multiparty_democracy>	$11/11 = 1$
one party	<wikipedia:Single-party_state>	$5/983 = 0.005$
one party	<wikipedia:Non-possessors>	$1/983 = 0.001$

Step 3: Disambiguation

CMNS performs disambiguation by returning the top ranked entity for each mention, when the ranking score is above the threshold τ_s.

Question. Considering $\tau_s = 0.01$, what is the output of the CMNS approach?

Mention	**Entity**
1992 elections	<wikipedia:Philippine_general_election,_1992>
angola	<wikipedia:Angola>
multiparty democracy	<wikipedia:multiparty_democracy>