Annotated bridging anaphora sample corpus for Russian

Anna Roytberg, PhD student, NRU HSE

Bridging: background and terminology

I looked into the **room.** The **ceiling** was very high. [Example from [Clark]

Bridging element – is an anaphoric element refers to the anchor (*room* is in the example)
Anchor - is the noun phrase to which an bridging-element refers (*ceiling* is in the example)
Bridging link – the link from bridging-element to anchor

We've started with bridging in genitive construction

Genitive construction:

N + (PRON gen) + (A gen) + N gen

Bridging in genitive construction

Bridging element: N_1 + ~~(PRON gen) + (A gen) + N_2 gen~~
Anchor: N_3 coreferented to N_2

[N_3] [N_1] [N_2]
В автобусе начался пожар. Водитель {~~автобуса~~} сам потушил огонь.
'In' 'bus' 'start'-PST 'fire' 'driver' {'~~bus'-Gen~~} REL-PR-3d-S 'put out' 'fire'
'The fire broke out in the bus. The driver put out the fire by himself'

Sample bridging corpus

We have annotated 250 short news.

(short = near to 100 words each)

What do we tag?

- Just bridging elements, anchors and the bridging-links.
- Later we want to tag POS, topics etc.

The Corpus statistics

Texts	250
Bridging links in all the corpus	160
Bridging clusters	100
The size of a cluster (words)	min=1, max=8, avg=1.58
The length of bridging link (words between entities)	min=0, max=299,avg=41

Some intresting cases

Bridging vs. coreference

Об этом сообщил [**адвокат** [**одного из участников** акции] $_{Gen}$]$_i$ Александр Мельников, сообщает РАПСИ.
lawyer *of one of the participant of the action*

По словам [**защитника**]$_i$, в эту пятницу уже шестеро британцев покинули Россию, один уедет в течение дня.
barrister

There are 2 possible variants

1. адвокат [одного из участников]$_i$ is corefereneted with [защитника]$_i$ - more preferable
2. [защитника] has a briging link to [одного из участников]

We tag the both variants : адвокат [одного из участников]$_i$ —coref→ [защитника]$_i$ + [защитника] —bridge-coref→ [одного из участников]
lawyer *of one of the participants* *barrister* *barrister* *of one of the participants*

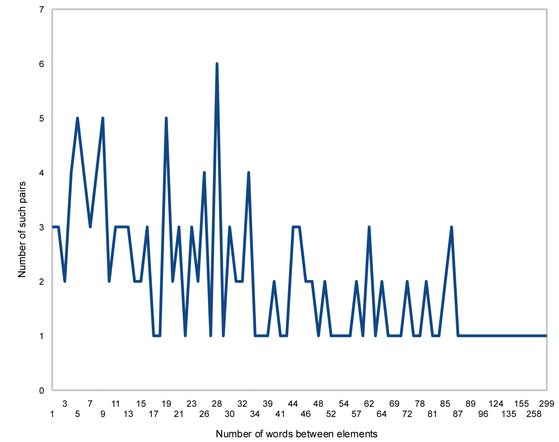
Hidden genitive bridging

Московский **таксист** отобрал у **пассажира** {~~такси~~}$_{Gen}$ 1,5 млн рублей.
taxi driver *passenger* *(of the taxi)*

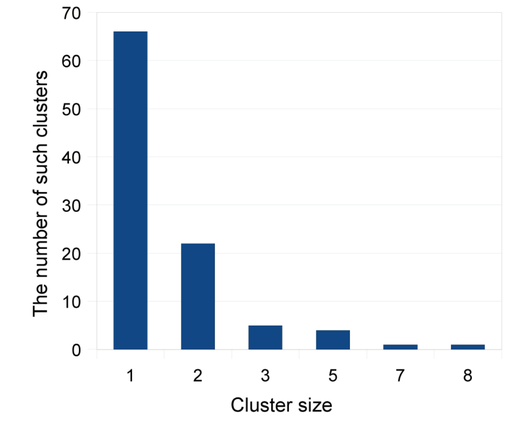
We have no word "taxi" in pretext. But it's "hidden" in "taxi driver'. NB In Russian we have no collocation "taxi driver', but the single word "таксист".
We tag: пассажира —hidden-bridge→ таксист

Preliminary data

1. The length between bridging-element and an anchor



2. Clusters size



3. There are very few bridging links in a law (or nearly law) texts

4. Toponims are the most popular anchors in genitive bridging (generally countries and cities names)

References

1.Clark, H. H. 1977. Bridging. In Johnson-Laird and Wason, eds. Thinking: Readings in Cognitive Science. Cambridge University Press, Cambridge.

2. Hou, Y.,Market,K., Strube, M. 2013 Cascading Collective Classification for Bridging Anaphora Recognition using a Rich Linguistic Feature Set. EMNLP 2013: 814-820

3. Poesio, M. and Vieira, R. 1998. A corpus- based investigation of definite description use. Computational Linguistics, 24(2):183–216.

4. Poesio, M., Mehta, R., Maroudas, A., Hitzeman,J. 2004. Learning to resolve bridging references. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004, pages 143–150.