ELDEN: Improved Entity Linking using Densified Knowledge Graphs - Supplementary Document

Priya Radhakrishnan
IIIT Hyderabad, India
priya.r@research.iiit.ac.in

Partha Talukdar
Indian Institute of Science
Bangalore, India
ppt@iisc.ac.in

Vasudeva Varma
IIIT Hyderabad, India
vv@iiit.ac.in

1 Discussion

In this supplementary document, we analyze the driving intuition and model choices of ELDEN and cases where it does not improve EL performance.

1.1 ELDEN Intuition :

Normalized Google Distance (NGD) (Cilibrasi and Vitanyi, 2007) is a semantic similarity measure derived from the number of hits returned by Google search engine for a given set of terms. It is based on term occurrences on web-pages. Web-pages that contain both terms indicate relatedness. WLM (ψ_{wlm}) is NGD adopted with edges. In ψ_{wlm}, an edges common to two entities in Wikipedia, indicate relatedness. We argue that ψ_{dense} is also NGD, as it is ψ_{wlm} extended with positive PMI edges. Here the additional edges are positive PMI entity(ies) which is common to both the entities and thus indicate relatedness. Further, in ψ_{wlm}, common edge is a Wikipedia page that contains both entities. In ψ_{dense}, pseudo entity is a term that co-occurs with both the entities.

1.2 Use of top K search results as web corpus:

To create the co-occurrence matrix M, ELDEN uses the top K search results. This requires $|E|$ search queries. The NGD method uses number of counts returned by Google, which requires $|E|^2$ search queries.

1.3 String matching to identify entity occurrences :

In the web corpus we used string matching to identify entity occurrences. This worked for sparsely-linked entities especially as ambiguity level was lower for sparsely-linked entities compared to densely-linked entities, as evident in Table 2.

edges	0-200	200-400	400-600	600-800
Ambiguity	12.55	38.55	32.66	87.23

Table 2: ambiguity level measured as (average of) number of unique entities a mention could link to, is lower for sparsely-linked entties compared to densely-linked entities.

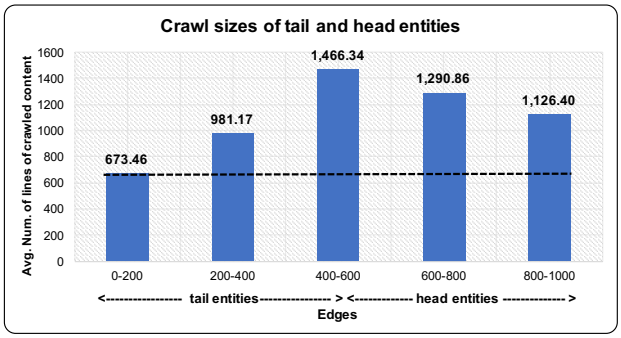

Figure 1: Sparsely-linked (tail) entities have average crawl size of 670 lines or more. Thus corpus proves to be a good source of additional links for densification, for both sparsely-linked and densely-linked entities.

1.4 Effect of training data size on Precision:

As ELDEN uses a web corpus for KB densification, we analyze the size of web corpus vis-a-vis precision in Figure 2. We see that precision does not vary greatly when size is more than 500 lines and we saw in Figure 1 that sparsely-linked entities have on an average more than 500 lines size.

1.5 PMI link availability :

One of the limitations in ELDEN's KB densification method is that, if the links from positive PMI entities to sparsely-linked entity is less or nil, it results in no densification of inlink graph. In some

Error Class	Example	Referent entity	Predicted entity
Acronym	I commute from Blue Ash (near Amberly) to the VA Medical Center (near **UC**). In the morning I shoot down Blue Ash Rd to Montgomery Rd.	University of Cincinnati	University of California
	Recent flooding in New South Wales put the pressure on the insurers, with **IAG** down 0.07 at 5.90 while competitors AMP slipped to..	Insurance Australia Group	International Airlines Group
Synonym	a federal study that cites rising tobacco use and higher costs for the Pentagon and Department of Veterans Affairs as reasons for the ban. The study by the Institute of Medicine, requested by the **VA** and Pentagon, calls for a phased-in ban over a period of years, perhaps up to 20.	United States Department of Veterans Affairs	Virginia
Specific label	each year, the oil firm would owe just over $195,000. estimated 4.4 million tons of carbon dioxide emitted (based on Others in the top 10 include the Chevron refinery in **Richmond**, a power plant in Pittsburg and a cement company in Cupertino.)	Richmond, California	Richmond, Virginia
Miscellaneous	**MILAN**, Italy has not been able to overcome Les Bleus since, losing 3-1 in Italy beat France on penalties to win last year's World Cup, but Paris and drawing 0-0 in Milan in qualifying for Euro 2008.	Milan (city in Italy)	MILAN (the football team from Milan in Italy)

Table 1: Categorizing error made by ELDEN. Examples from TAC dataset. Better modelling of contextual entities and coherence of contextual entities can help reduce 'Specific label' errors (Please see Sec 1.6)

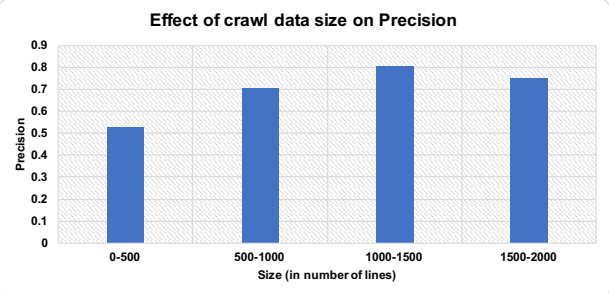

Figure 2: Average size (number of lines) of crawled content of an entity versus precision in EL.

cases new edges are added by PMI, but the added edges are not common between the pair of entities. This also results in no densification.

1.6 Error Analysis of ELDEN

We analyzed errors made by ELDEN on TAC dataset, before and after applying KG densification features. We categorize the errors into four classes (Table 1) , in line with error classes of Ling *et al.* (2015). Table 1 shows the mapping. An error is in "Synonym" class when mention is an alternate name of referent entity as in '*Tamil Tigers*' for 'Liberation_Tigers_of_Tamil_Eelam'. [1] and "Specific label" class if referent entity is a specific instance of the mention. . For example, the referent entity could be 'Westfield,_NJ', 'Westfield_Group', 'Westfield_Corporation' and so on for the mention '*Westfield*'. An error is in "Acronym" class when mention is an acronym and referent entity is expansion of the acronym like '*NSA*' and 'National_Security_Agency'. Errors in (The error here could be due to wrong co-reference resolution or wrong coherence) , where the mention is not detected to be close to other similar (location or organization) entities, or it could be due to wrong type identification of mention. Thus the classes 'Co-reference', 'Coherence' and 'Type' of Ling *et al* map to our "Specific label" class. All other errors are accounted in "Miscellaneous" class.

[1] this class roughly translates to Ling *et al.*ś Metonymy class

1.7 Source Code and Data :

The source code is available at https://github.com/priyaradhakrishnan0/ELDEN. README contains the necessary execution instructions. The repository contains code for training entity-embeddings and PMI co-occurance matrix. The link to download the pre-trained entity embeddings file and PMI co-occurance matrix is also present in the repository.

References

Rudi L. Cilibrasi and Paul M. B. Vitanyi. 2007. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.* 19(3):370–383. https://doi.org/10.1109/TKDE.2007.48.

Xiao Ling, Sameer Singh, and Dan Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics (TACL)* 3.