Generating a Large-Scale Entity Linking Dictionary from Wikipedia Link Structure and Article Text

Ravindra Harige, Paul Buitelaar
Insight Centre for Data Analytics
National University of Ireland, Galway
[firstname.lastname]@insight-centre.org

Abstract

Wikipedia has been increasingly used as a knowledge base for open-domain Named Entity Linking and Disambiguation. In this task, a dictionary with entity surface forms plays an important role in finding a set of candidate entities for the mentions in text. Existing dictionaries mostly rely on the Wikipedia link structure, like anchor texts, redirect links and disambiguation links. In this paper, we introduce a dictionary for Entity Linking that includes name variations extracted from Wikipedia article text, in addition to name variations derived from the Wikipedia link structure. With this approach, we show an increase in the coverage of entities and their mentions in the dictionary in comparison to other Wikipedia based dictionaries.

Keywords: Entity Linking, Wikipedia, DBpedia, Name Variation extraction

1. Introduction

Entity Linking is the task of linking mentions in text to entities in a knowledge-base (KB) like Wikipedia and assign it an unambiguous entity (sense) identifier. This task is challenging and comprises of two main steps: i) identifying phrases to annotate and ii) finding and linking the right entity in the KB, considering its context. These two steps are called phrase detection and disambiguation. In this work, we mainly focus on one specific subtask of the disambiguation step, namely candidate generation.

Typically, a pre-computed dictionary derived from the KB is used for candidate generation as we do here with Wikipedia as a KB. The quality and scope of this dictionary plays an important role in the disambiguation process as its performance is not only dependent on the actual algorithm used to disambiguate but also on the quality and expressiveness of the dictionary used to determine the entity candidates for the phrase mentions in text.

We use patterns to analyse the leading sentences of Wikipedia articles and query structured infobox information captured by DBpedia to gather name variations for entities, extended it with name variations mined from the Wikipedia link structure. Additionally, we employ a method to consolidate name variations extracted from different approaches into one set of mappings, which is essentially an entity linking dictionary. We provide a two-way mapping, such that an entity in the KB can be reached from an entity mention and all entity mentions can be queried for any given named entity in the KB.

2. Related Work

Several approaches exist for generating an entity-linking dictionary from Wikipedia or other knowledge resources. We review some of the well known entity linking dictionaries in the following.

Google released a cross-lingual dictionary for English Wikipedia concepts consisting of 175 million surface forms that refer to 7.6 million entities derived from Wikipedia article titles, anchor text from inter-Wikipedia links, non-Wikipedia web pages linking to Wikipedia articles and non-Wikipedia web pages to non-Wikipedia pages for topics that have corresponding Wikipedia articles(Spitkovsky and Chang, 2012).

DBpedia lexicalization dataset was released as a part of DBpedia Spotlight project. It contains alternative names for entities and concepts in the DBpedia project with several scores estimating the association strength between name and URI(Mendes et al., 2012).

Similar to DBpedia Spotlight, the AIDA project developed a dictionary called 'YAGO means' for use in entity disambiguation and linking. It is constructed by extracting link anchors, re-direction and disambiguation page links in Wikipedia (Yosef et al., 2011). A slight variant of YAGO Means, Redirect Disambiguation Mapping (RDM) dictionary contains the additional entries with alternative labels of DBpedia entities. For example, the label "Berlin (2009 film)" has been converted to just "Berlin" (Steinmetz et al., 2013).

3. Wikipedia as a Lexical Resource

To construct a dictionary from Wikipedia, we first look at the various data sources that could be mined for extracting name variations. We find three data sources of interest for this purpose: 1) Wikipedia articles data (unstructured text) 2) structured information from infoboxes (DBpedia); and 3) disambiguation and redirection links. As the existing literature have already detailed on resourcefulness of Wikipedia's internal links, disambiguation and redirection links; in the following we illustrate how Wikipedia's article text and structured information from infobox are also an interesting source for extracting entity name variations.

In Wikipedia article text data, we are primarily interested in two parts: the lead sentence of the article, henceforth referred to as Topic Sentence (TS), the body of the article text.

Topic Sentence (TS): The TS defines the topic and introduces other common or popular names for the topic. As a TS is typically the start of or the synonym for the article

Chen Yucheng	
Nickname(s)	Four-eyed dog
Born	1837 Teng County, Guangxi, Qing Empire
Died	1 May 1862 (aged 24–25) near Xinxiang, Henan
Allegiance	Qing Empire (to 1849) Taiping Heavenly Kingdom (to 1862)

Figure 1: Infobox consisting nickname (name variariation)

title, and goes on to introduce other names for the article, it is ideal to extract name variations from it. Consider the following TS examples in which name variations are highlighted in bold-italics:

Melon-headed whale. The melon-headed whale (species Peponocephala electra; other names are ***many-toothed blackfish*** and ***electra dolphin***) is a cetacean of the oceanic dolphin family (Delphinidae).

Isoorientin. Isoorientin (or ***homoorientin***) is a flavone, a chemical flavonoid-like compound. It is the luteolin-6-C-glucoside. Bioassay-directed fractionation techniques led to isolation of isoorientin as the main hypoglycaemic component in Gentiana olivieri.

Joris Jarsky. Joris Jarsky (born December 3, 1974), also known as ***Joris Jorsky***, is a Canadian stage, film and television actor who has received recognition for being a versatile actor, and is known for his role as Marty Strickland in the series Vampire High.

Article body: Throughout the article body there are often mentions of acronyms with their expansions in the form of NAMED ENTITY (ACRONYM). If the acronym expansion phrase is hyperlinked, it can be considered as the name variation for the Wikipedia article that it links to.

Structured information from the infobox: Some Wikipedia articles mention alternate names like nickname(s), stage name(s), etc. in their infobox. For example, the article on Chen Yucheng, a Chinese general lists his nickname as "Four-eyed dog" (see Figure 1). This information is mentioned in a template/structured form.

4. Extraction Approaches

We perform different experiments to extract name variations from each of the different parts of Wikipedia. The experiments are grouped in five categories as shown in the Table 1, each of which produces name variations extracted by the methods in corresponding categories.

Among the listed categories, WTS and WAA are particularly challenging, as they involve extracting name variations from free text. In the following, we explain each of our extraction approaches in some more detail and also discuss our method for consolidating the extraction results across the five methods into a comprehensive entity linking dictionary.

Table 1: Category of experiments performed to extract name variations

Category	Data derived from	Name
WTS	Article Text: Topic Sentence	D1
WAA	Article Text: Acronyms and Abbreviations	D2
WIL	Article Text: Article Title and Inter-article links	D3
LDQ	Infobox: Linked Data RDF	D4
RDL	Redirect and Disambiguation links	D5

4.1. WTS: Extracting name variations from a Wikipedia Topic Sentence

To extract the name variations from the Topic Sentences, we first analysed the lead paragraphs of Wikipedia articles with the objective to: a) identify what are the common conventions or linguistic styles used by editors of Wikipedia articles to mention name variations in the TS; b) see if these conventions can be generalised into a set of patterns that define ⟨TOPIC⟩ to ⟨Name Variation⟩ relations.

We identified 11 broad types of patterns in which editors of Wikipedia articles mention the name variations in topic sentence. For each of the patterns, 50 TS were selected and annotated with the name variations in it to create a ground truth dataset. Using this, we trained a linear chain Conditional Random Field model to label and extract the name variations in the topic sentences.

4.2. WAA: Extracting acronyms and abbreviations from a Wikipedia article

In this method, the objective is to identify abbreviations and their expansions in the free text of a Wikipedia and extract the corresponding Wikipedia topic URI for it, if it exists as the hyperlink for either acronym or the expansion string. We defined a rule based system using regular expressions to identify abbreviations and their expansions in Wikipedia articles and extract the corresponding Wikipedia topic URI for it, if it exists.

4.3. WIL: Mining Wikipedia inter article links for name variations

Wikipedia articles are strongly connected to each other by hyperlinks. The important phrases or mentions in the article that can be further explored are usually hyperlinked; such mentions are called *Anchor Text*. We defined a regular expression based parser to extract hyperlinks and corresponding anchor texts from the inter-article links.

4.4. LDQ: Querying name variations from structured information in Wikipedia

As seen in previous section, the name variations found in infobox are in the structured form, making it easier to manually extract the correct name variations for any given topic. However, it is well known that the goal of DBpedia (Mendes et al., 2012) is to extract structured information from Wikipedia and make it available on the web for querying. Thus we decided to use DBpedia data for querying the name variations from it.

4.5. RDL: Mining name variations from Redirects and Disambiguation links

A disambiguation page title is a valid name variation for all the Wikipedia articles that the page points to. Redirect URLs, although they do not have a page for themselves like for Disambiguation pages, can be considered a valid name variation for the article it redirects to. It is possible that a redirect link goes to a disambiguation page, for this reason, we consider the source string which triggered the redirect as the name variation for all the entities that disambiguation page points to.

4.6. Consolidating Name Variations

From each of the extraction method categories, we have extracted name variations in the form of tuples such that each Wikipedia Entity has one or more unique name variations, with the number of their occurrences in the dataset Di. It is very likely that some of the name variations extracted from different approaches are repeated. Thus in this task, we integrate extractions from each dataset into one dictionary and assign a score indicating strength of the association between E and V quantified with probability statistics.

$$D_i = (E_j, V, count(E_j, V))$$

Where,
$D_i = Dataset_i; i \in \{1..5\}$,
E_j = Unique Wikipedia Entity URI; where, $0 \leq j \leq$ total no. of unique entities in D_i,
V = Unique Name Variation for E_i, and
$count(E, V)$ = count of the pair (E_j, V) occuring in D_i

5. Entity Linking Dictionary

5.1. Data modelling

The final, consolidated version of dictionary dataset is modelled in RDF format using SKOS- XL (Mile and Bechhofer, 2009) schema. The other candidate RDF schemas was Lemon (McCrae et al., 2012), the purpose of which is to enable people to be able to share lexicons on the Semantic Web. However, to model a dictionary in Lemon, it is necessary to identify name variations by morphology, spelling variants, etc. As our data is automatically extracted, the only distinction we make is in written representation; the morphology and spelling variations are not readily available in our data. For this reason, we chose the SKOS-XL model which is less linguistic in nature and does not have this requirement.

5.2. Evaluation

We chose two gold standard datasets, the DBpedia Spotlight NER corpus (Mendes et al., 2012) and the KORE50 data set (Hoffart et al., 2012) from the AIDA project, released for Named Entity Disambiguation as the benchmark to evaluate our generated dictionary. The evaluation approach is adopted from (Steinmetz et al., 2013), however, we choose only Maximum Recall metric to evaluate the recall of our dictionary (OD) on DBpedia Spotlight and the KORE50 data set, and compare it to the reported Maximum Recall of other dictionary datasets such as DBpedia Lexicalization (DBL), Redirect Disambiguation Mapping (RDM), AIDA means and Google Cross Wiki (GCW). We report the number of entries in our dictionary and the recall of our dictionary in comparison to other dictionaries of (Steinmetz et al., 2013), in Table 2 and Table 3. The DBpedia Spotlight NER corpus consist of annotations for 60 sentences from 10 New York Times articles. The annotation results in 249 DBpedia entities, whereas KORE50 consist of annotations for 50 sentences from different domains like music, business and celebrities, and is based on the CoNLL 2003 NER task data set. Distinct mentions of all annotated entities in each of these are looked up in the dictionary if it exists there. The number of mentions found in our dictionary is the maximum recall.

Table 2: Mention counts in different dictionary datasets

Dictionary	DBL	RDM	AIDA	GCW	OD
No. of entries	2M	10M	18M	378M	23M

Table 3: Maximum Recall in different dictionary datasets

Dictionary	DBL	RDM	AIDA	GCW	OD
Spotlight (265 mentions)	85% (224)	86% (228)	28% (75)	91% (242)	88% (233)
KORE 50 (130 mentions)	68% (89)	72% (93)	86% (112)	85% (110)	80% (104)

Our dictionary contains names of person, places, organizations, etc. and thus has better coverage of entities compared to DBL and RDM datasets. For the same reason, maximum recall is better than both DBL and RDM on both the benchmark datasets, but the AIDA dictionary has slightly better recall on KORE50 than ours. Overall, our dictionary has slightly better recall than all other dictionary datasets except Google Cross Wiki. GCW is a comprehensive dictionary as it is prepared by leveraging web crawl data that is not practical to prepare in an academic setting.

6. Conclusion

We explored different parts of Wikipedia to extract name variations for the entities in Wikipedia, and consolidated all the name variations into one single dictionary. Our dictionary generation method provides the mappings to reach an entity from a text mention string and also to retrieve all the name variations for the entity. The initial evaluation results show that the name variations extracted from Wikipedia article text, in addition to the Wikipedia link structure, do contribute to overall increase in the coverage and recall of entities when resolving a text mention to a Wikpedia entity. We see two major outcomes of our work as follows: i) the generated dictionary for use by the research community in Entity Linking and other NLP related tasks; and ii) the framework software to extract name variations from Wikipedia which will enable anyone to extract a dictionary on newer versions of Wikipedia data[1].

[1]Dataset and extraction framework code will be made available at https://github.com/ravindraharige/lrec2016

7. References

Hoffart, J., Seufert, S., Nguyen, D. B., Theobald, M., and Weikum, G. (2012). Kore: keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 545–554. ACM.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., et al. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

Mendes, P. N., Jakob, M., and Bizer, C. (2012). Dbpedia: A multilingual cross-domain knowledge base. In *LREC*, pages 1813–1817.

Mile, A. and Bechhofer, S. (2009). Skos simple knowledge organization system extension for labels (skos-xl) namespace document - http://www.w3.org/tr/skos-reference/skos-xl.html.

Spitkovsky, V. I. and Chang, A. X. (2012). A cross-lingual dictionary for english wikipedia concepts. In *LREC*, pages 3168–3175.

Steinmetz, N., Knuth, M., and Sack, H. (2013). Statistical analyses of named entity disambiguation benchmarks. In *NLP-DBPEDIA@ ISWC*.

Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011). Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453.