Entity Linking through Neighborhood Comparison and Random Walks*

Chuan Liu
Johns Hopkins University
3400 N. Charles Street
Baltimore, Maryland 21218
chuan@cs.jhu.edu

Abstract

In this project, we explore two kinds of methods for the entity linking task. The first kind compares the similarity between entities by their common neighbors, and second is based on random walk models. Experimental results on data extracted from Wikipedia are shown for both kinds of methods.

1 Introduction

The entity linking task in Knowledge Base Population Track of Text Analysis Conference 2009 (McNamee, 2009) is introduced as a problem of "name-string disambiguating". It can be viewed as an unsupervised name entity disambiguation problem at large scale. Given a list of entities, each entity may have one or multiple surface names and a surface name may refer to different entities. For each entity, there is an associated article or descriptive document. From the article, we can extract a list of surface names of related entities. However since different entities can share a same confusable name, we may not know what exact entity the surface name refers to. The task is to uncover the links between the surface names and the true entity. The concepts can be further illustrated with the following example.

> Michelle LaVaughn Robinson Obama (born January 17, 1964) is the wife of the forty-fourth President of the United States, Barack Obama, and is the first African-American First Lady of the United States.
>
> ... She is the mother of two daughters, Malia and Sasha, and is the sister of Craig Robinson, men's basketball coach at Oregon State University.

In the example, the personal entity is named "Michelle Obama"[1], followed is a short introduction of her extracted from Wikipedia (Wikipedia, 2009a). The surface names in the piece of text are underlined. For some name, like "Barack Obama", it easy to identify the entity. For others, like "Malia" and "Sasha", it is difficult given only the surface names. Even for the less common name "Craig Robinson", we have following potential entities from Wikipedia.

- Craig Robinson (actor) (b. 1971), actor and stand-up comedian
- Craig Robinson (baseball) (fl. 1970s–80s), Major League infielder
- Craig Robinson (basketball) (b. 1962), Oregon State basketball coach
- Craig Robinson (designer) (b. 1972), fashion designer
- Craig Robinson (rugby league) (b. 1986), rugby league player

*This is the project report for the Machine Learning course taken at Johns Hopkins University in the 2009 Fall.

[1]We will refers to the person through its names when it is clear out of context, i.e. no ambiguity, like in this example, it is clear who we are referring to with the short introduction. In other cases, we may not know which entity the name refers to, and that is what we want to resolve in this task.

We only considered personal entities in the above example. In reality, we may as well have other types of entities, including geopolitical entities like "United States", organizational entities like "Johns Hopkins university", and so on. In the even worse scenario, the surface name may suggest little or give misleading clue about the underlying entity, e.g. when we refers to the jazz musician "George Washington", the name may mislead us to think of the president of the same name. The diversity of entities and their names makes the task more complex and difficult.

2 Methods

Like all unsupervised learning algorithms, the key step is to design a good measure of distance between instances. With a good measure, the algorithm simply choose the closest entity, i.e. the entity with shortest distance, as true entity to the given surface name.

In this project, we investigate two types of measures. The first kind compares entities directly based on their related surface names; the second try to explore the structure information between entities by utilizing the random walk model. If we think of the related surfaces names of the entity as neighbors of the instance, the first type of methods, which measure similarity by directly comparing the related surface names of the two entities, can be viewed as neighborhood comparison. Alternatively, take the neighborhood relation as an edge between two entities. On the whole, the entities form a large network or graph. The second tries to capture the structural information in the large network. We give a formal introduction to our methods in the rest of the section.

Suppose we have some measure of distance between entities x and y denoted by $d(x, y)$. Given an entity x and surface name $\bar{y}$ and let $\{y_i\}$ be potential target entities of $\bar{y}$. We choose $y = y_j$, where $j = \arg\min_i d(x, y_i)$, as the true entity $\bar{y}$ refers to. The idea is that if y is mentioned in x's description, they should be related. Thus we choose the closest entity in terms of the measure d as the true entity.

2.1 Neighborhood Comparison

This line of methods is motivated by research in and social network research (Liben-Nowell and Kleinberg, 2003). Given two entities, we can measure their similarity based on their neighborhoods. Here neighborhoods are defined as the entity surface names found in the description article. We give three common used measures of this type below. For entity x, $\Gamma(x)$ denotes the set of x's neighbors.

Common neighbors defines

$$d(x, y) := |\Gamma(x) \cap \Gamma(y)|,$$

the number of common neighbors of entity x and y.

Jaccard's coefficient defines

$$d(x, y) := \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

measures the number of common neighbors both x and y have compared to neighbors either x or y have.

Adamic/Adar similarity defines

$$d(x, y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|}$$

where $|\Gamma(z)|$ is frequency of z. The measure gives more weights to rarely appeared entity in neighbors.

The idea here is the relation between x and y can be testified by the neighbors they have in common as analogue to comparing similarities between persons in the social network.

2.2 Random Walk Model

First, we define the entity graph G as follows. Each entity is a vertex in the graph. Every related surface name $\bar{y}$ for a entity, say x, may have one or several potential target entities $\{y_i\}$. An edge in the graph is a pair of (x, y_i).

In the random walk model, we assume we can travel from a vertex to another if there is an edge between them. The (one step) transition probability P for the graph is defined as a $N \times N$ matrix, where N is number of vertices and $P_{x,y}$ is the probability of traveling to y in next step given we are at vertex x. The probability we can travel from x to y in n steps in matrix form is then

$$P^{(1)} = P$$
$$P^{(n)} = P^{(n-1)} \cdot P = P^n$$

Then the neighborhood random walk distance $d_n(x, y)$ is defined as

$$d_n(x, y) = \frac{\sum_{k=1}^{n} P_{x,y}^{(k)}}{n} \quad (1)$$

where n is steps of random walk. In words, $d_n(x, y)$ is the normalized total probability of reaching y in $1, ..., n$ steps if we start the random walk at x.

3 Data Collection

Data collection took a large portion of our effort in this project. It is also important to understand on what data the experiments are carried out. In this section, we describe the collected data.

The data are collected from English version of Wikipedia (Wikipedia, 2009b). For each entity, the links in the article are extracted. The links in Wikipedia markup language look like the following examples.

```
[[Faustina|Faustina_the_Elder]]
[[Otto I|Otto_I,_Holy_Roman_Emperor]]
[[Paul I|Paul_I_of_Russia]]
```

where the text before "|" is the displayed name for the link, and the text after the separator is the canonical form of the link. If the two are the same (as in most cases), the later part can be omitted. In general, each entity has an unique canonical name. We will use that as an identification to the entity. For each entity, the link text will be its related surface names. We also store the true target for each link. This information is only used for evaluation.

Another source of information we extracted from Wikipedia is from the disambiguation pages. For many ambiguous names, Wikipedia has a disambiguation page that list all the potential candidates for that name. For example, the disambiguation page for the name "George Washington" has the following potential candidate entities.

- George Corbin Washington
- George Washington (inventor)
- George Washington (Washington pioneer)
- George Washington (trombonist)
- George T. Washington (Liberia)
- George Thomas Washington
- George Washington (baseball)

For each of such pages, we extract the name and the list of potential candidates. We also want to note, like all other Wikipedia pages, these pages are also edited by volunteers. As such the list of candidates may not necessarily be complete and are subject to errors. This issue will be addressed further in Section 4 when discussing experiments.

4 Experiments

This section discusses the data used for experiments (Section 4.1), experiment settings and the experimental results (Section 4.2).

4.1 Data Set

The collected data from Wikipedia is of huge amount. We restrict our attention to a small subset of data that is feasible within the scope of this project. We briefly explain the restrictions as follows.

First, we only try to solve the ambiguity for the surface names appeared as disambiguation pages. With this restriction, we also have a good list of potential candidates given by the disambiguation page. We will look for the true entity among the suggested candidates only.

Second, as noted before, the list of candidates given in the disambiguation page are not necessarily complete. Indeed, there are quite a few links, whose target does not exist in the candidates given by the disambiguation page. We omitted all such links in the experiments, for otherwise we cannot find the true target by the first restriction anyway. Note this essentially gives us a 100% recall in experiments.

Third, total links in the article are still too huge a number. When we ran the code on the whole set of links, the program rapidly used up 2GB memory of the computer. In order to get results in a reasonable short time, we restrict to the links that appears in the infobox only. An infobox (as shown in Figure 1) is the side box appeared on top right corner of a Wikipedia page shows some summary of facts in a structured way for some entities. Note not all entity has an infobox, and the number of links in the infobox is much smaller than that in the article.

Figure 1: Part of an infobox for "George Washington" extracted from Wikipedia. The blue text are all hyper links. We need to find the true entity given the link text.

Total entities	5856
Average links per entity	2.83
Average ambiguity	4.36

Table 1: Some statistics of the data set.

In conclusion, we include the following two kinds entities (together with their surface names) in our experiments, and some simple statistics about the data set are listed in Table 1.

1. Entity that contains at least one ambiguous surface name, i.e. the name appears as a disambiguation page title, in its infobox link.

2. All the candidate entities that are given by the disambiguation pages.

4.2 Method Evaluation and Results

All the previously described methods, as well as two more simple methods (including one as baseline method), are evaluated on the data describe in previous section. Also as noted in previous section, we have 100% recall, so the evaluation is solely based on accuracy which is calculated as

$$\text{accuracy} = \frac{\text{number of correct entities resolved}}{\text{total number of names to resolve}}$$

The baseline method always selects the entity with most number of surface names among the potential candidates in the sense it is most important or the most popular choice as it draws most attentions from volunteers to create links for it.

Exact match always selects the exact match first. If no exact match, back off to baseline method.

Random walk model For the random walk based method, we tried $n = 1, 2, 3, 4, 5$, in hope to have an idea how the change of n affects the model.

Neighborhood Comparison All the three measures described before are implemented. When there is a tie in distance, we always back off to the baseline method.

The results are shown in Table 2. Overall, random walk based method outperforms neighborhood comparison methods. Among random walk based methods, when $n = 1$, we have the best result. The common neighbors measure gives best result among neighborhood comparison methods.

All the methods shows better performance than the baseline approach and worse performance than the exact match. There are more exact matching links in the data since when people create links, they tend to prefer a simpler form (omitting the target part after "|" as mentioned before). It is then of interests to try out both types of methods on none exact match links for comparison purpose. So we run exact match and let it back off to random walk and neighborhood comparison respectively. The result is 98.4% for both methods (listed as "exact match*" in the Table 2). Examining more figures in decimal, it turns out random walk actually performs slightly better, but the difference is less than 0.1%. However by further inspecting the results, it can be noticed the errors of the two methods are different, however both kinds of errors are rather random. We list some sample errors made by the two type of methods exclusively in the Table 3 at the end of the report.

The data favors simple models for both types of methods. An observation is that our data is very simple and restricted. This may limit the capability of

Method	Setting	Accuracy
Baseline		84.4%
Exact match		98.3%
Neighborhood Comparison	Common Neighbors	86.1%
	Jaccard's coefficients	85.3%
	Adamic/Adar similarity	85.2%
Random Walks	$n = 1$	92.9%
	$n = 2$	91.9%
	$n = 3$	86.9%
	$n = 4$	86.9%
	$n = 5$	86.9%
Exact match*		98.4%

Table 2: The experimental results. The result for "exact match*" as explained before is the same result for several variations of exact match method.

the model. Experiments may be carried out to further to investigate the performance of the models on larger and more complex data. It is also worth noting even on such simple data, random walk based methods are more accurate than the neighborhood comparison.

5 Related Work

For the problem, the work by (Adafre and de Rijke, 2005) is related to ours. They are trying to find missing links on Wikipedia pages. The difference here is we are more concern about finding the true page given the link text while they are more focus on detecting potential link text in the given description. We also note since their source of data are also from Wikipedia, there are also similarities in data extraction and collection between the two works.

For the distance measures, (Liben-Nowell and Kleinberg, 2003) did extensive studies on measuring distance in large and complex graph for the link prediction problem in social networks. (Bhattacharya and Getoor, 2007) used neighborhood similarity measure in the task of entity resolution. Enormous literature exist for the random walk model. The most famous may be the work of PageRank (Brin and Page, 1998).

6 Conclusion and Comparison to Proposal

In conclusion, we have implemented all the proposed methods for the entity linking project with one exception that we mentioned a recent work by (Zhou et al., 2009) in the proposal as an natural extension to our work. However much time is spend on tuning data to get the program work in later days, and the method is not implemented.

Though not mentioned explicitly in the proposal, we intended in mind to carry out experiments with much larger data sets (using article links). However after tried several ways pruning the data, it still not works out, and we have to switch back to the currently used much smaller data set to meet the deadline. The experiments on large scale data are left for further investigation.

The code and data are available for download at http://www.cs.jhu.edu/ chuan/cs475_project.tar.bz2.

References

[Adafre and de Rijke2005] Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering missing links in wikipdia. In *Proceedings of the 3rd International Workshop on Link discovery*.

[Bhattacharya and Getoor2007] Indrajit Bhattacharya and Lise Getoor. 2007. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*.

[Brin and Page1998] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*.

[Liben-Nowell and Kleinberg2003] David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*.

[McNamee2009] Paul McNamee. 2009. Text analysis conference 2009 knowledge base population track. http://apl.jhu.edu/ paulmac/kbp.html.

[Wikipedia2009a] Wikipedia. 2009a. Michellel obama. http://en.wikipedia.org/wiki/Michellel Obama.

[Wikipedia2009b] Wikipedia. 2009b. Wikipedia. http://www.wikipedia.org/.

[Zhou et al.2009] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. 2009. Graph clustering based on structural/attribute similaries. In *Proceedings of the 35th International Conference on Very Large Data Bases*.

Table 3: Some sample errors made exclusively by random walk based method ($k = 1$) and neighborhood comparison method (using common neighbor measure) respectively.

Entity	**Surface name / True Target**	**Error output**	**Potential candidates**
Saint Mercurius	Saint Basil / Basil of Caesarea	Basil the Elder	Basil the Elder, Basil of Caesarea, Basil the Confessor, Basil Fool for Christ, Basil of Ostrog
Grand Duchy of Baden	Napoleon / Napoleon I of France	Napolean (actor)	Napoleon I of France, Napoleon II of France, Napoleon III of France, Louis Bonaparte, Napoleon Louis Bonaparte, Joseph Bonaparte, Napoleon B. Broward, Napoleon Bonaparte Brown, Napoleon Chagnon, Napoleon Collins, Napoleon J.T. Dana, Napoleon Einstein, Napoleon Hill, Nap Lajoie, Napoleone Orsini, Napoleon Perdis, Napoleon XIV, Napoleon Zervas, Napolean (actor), Hisaye Yamamoto, Napoleon (rapper), Napolean (actor)
Adlai Stevenson	John Kennedy / John F. Kennedy	John Alexander Kennedy	John Alexander Kennedy, John F. Kennedy, John F. Kennedy, Jr., John L. Kennedy, John N. Kennedy, John P. Kennedy, John Pitt Kennedy, John Stewart Kennedy, John Thomas Kennedy, Vikram (actor), John Kennedy (disc jockey), John Kennedy (engineer), John Kennedy (lawyer), John Kennedy (musician), John Kennedy (puppeteer), John Kennedy (theologian), John Kennedy, Jr. (footballer), John Kennedy (NASCAR), John Kennedy (Scottish footballer)
Errors made by random walk model only.			
Rudolf Steiner	Schiller / Friedrich Schiller	Karl Schiller	Friedrich Schiller, Christian Schiller, Eric Schiller, Ferdinand Canning Scott Schiller, Heinz Schiller, Herbert Schiller, Julius Schiller, Karl Schiller, Leon Schiller, Mayer Schiller, Philip W. Schiller, Solomon Marcus Schiller-Szinessy
Yuri Shargin	Engels / Friedrich Engels	Mary Tate Engels	Friedrich Engels, Ludwig Engels, Wera Engels, Stefan Engels, Robert Engels, Mary Tate Engels, Floortje Engels
William Sorell	George Arthur / Sir George Arthur, 1st Baronet	George K. Arthur	Sir George Arthur, 1st Baronet, George K. Arthur
Patti Austin	Gershwin / George Gershwin	Ira Gershwin	George Gershwin, Ira Gershwin, Frances Gershwin, Arthur Gershwin
Errors made by neighborhood comparison only.			