With a Little Help from my Neighbors: Person Name Linking Using the Wikipedia Social Network

Johanna Geiß
Heidelberg University
INF 348, 69120 Heidelberg, Germany
geiss@informatik.uni-heidelberg.de

Michael Gertz
Heidelberg University
INF 348, 69120 Heidelberg, Germany
gertz@informatik.uni-heidelberg.de

ABSTRACT

Driven by the popularity of social networks, there has been an increasing interest in employing such networks in the context of named entity linking. In this paper, we present a novel approach to person name disambiguation and linking that uses a large-scale social network extracted from the English Wikipedia. First, possible candidate matches for an ambiguous person name are determined. With each candidate match, a network substructure is associated. Based on the similarity between these network substructures and the latent network of an ambiguous person name in a document, we propose an efficient ranking method to resolve the ambiguity. We demonstrate the effectiveness of our approach, resulting in an overall precision of over 96% for disambiguating person names and linking them to real world entities.

Keywords

NEL; Social Network; Wikipedia; Wikidata

1. INTRODUCTION

There is an increasing need for approaches in information retrieval and Web mining to uniquely identify real-world persons based on mentions in text documents. For example, to obtain a textual summary or chronology for a person, respective mentions of that person in diverse types of documents need to be identified. Key to these approaches are person name disambiguation and linking techniques that identify the corresponding real-world entity for a person name. This is a hard problem in particular if for an underspecified name multiple possible matches exist. Among the numerous approaches that have been proposed, clustering approaches are popular and widely used [13, 16, 22]. Some use context features such as links or keywords [14, 16]. Many of these approaches either consider only a small set of documents and names or do not link a cluster to a real-world entity. Other approaches make use of knowledge resources like Wikipedia [3, 5, 9, 17]. A representative among these approaches that is becoming increasingly popular is the use of network structures [15, 18]. What is missing is an approach that employs 1) a large-scale repository of real-world persons, and 2) a social network structure among these persons in support of disambiguating person names in a wide range of documents that is efficient and easily extendable.

In this paper, we present a novel approach to person name linking using social network structures, namely the Wikipedia Social Network [7]. It consists of about 800K unique persons and 67.5M weighted relationship edges, which are determined based on co-occurrences of uniquely identifiable person names in documents. The proposed method works as follows. Given a text document containing ambiguous person names, there might be several possible matching candidate persons in the network. What distinguishes these candidates from each other is their neighborhood in the network. To determine which candidate provides the best match for the ambiguous person mention, we consider the neighborhood of the ambiguous mention in the document in terms of uniquely identifiable person names. This is done efficiently by exploiting information from the social network. Compared to approaches that rely on clustering, our approach links ambiguous mentions to real-world persons. Based on a subset of about 41M person mentions identified by the Stanford NE recognizer [6] in the English Wikipedia, we demonstrate the effectiveness of our approach, resulting in an overall precision of over 96%. The approach can be applied to any type of document. To compare our approach to state-of-the-art methods we used the AIDA CoNLL-YAGO dataset [11] and received an accuracy of over 84%, which is comparable to other approaches. Our approach even outperforms comparable systems with a precision of 94%.

We believe that large-scale social network structures in which real persons are associated with nodes provide a key ingredient to scalable and extensive name disambiguation methods. In particular, combining social networks with name disambiguation methods provides an effective means to enrich these networks as knowledge backbone for a variety of IR and text analysis tasks.

The remainder of the paper is structured as follows. After a brief review of related work in Section 2, we give an overview of the Wikipedia Social Network construction in Section 3. In Section 4, we detail our disambiguation model. The result on the Wikipedia test corpus and an extensive discussion of influencing characteristics are given in Section 5. In Section 6, we present the evaluation results on a widely used dataset, followed by a summary and discussion of ongoing work in Section 7.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891109 .

2. RELATED WORK

Many approaches for named entity disambiguation cluster person mentions (and the documents they occur in) referring to the same entity [1, 13, 14, 22]. These approaches rely on co-occurrences of person mentions and corresponding features from local context and/or from external sources. Different mentions are mapped to a single entity, but a link to real world objects is missing. Also, most approaches are tailored towards only small document and person sets.

The approaches in [3, 5] exploit the category and link structure in Wikipedia and collect contextual clues and category information for each entity to disambiguate and link them to a given list of entities. Semantic networks or graphs extracted from Wikipedia are used in [9] to compare mentions to concepts/topics. A unified graph-based approach to entity linking and word sense disambiguation using BabelNet is presented in [19]. Other methods make extensive use of topic models. Probabilistic inference within a topic model where each topic corresponds to a Wikipedia article is used in [12]. The focus of these approaches lies on information that describes a person entity, but the co-occurring person entities are not taken into account.

Most approaches that use network structures employ some kind of graph traversal and clustering [15, 18, 20]. The framework in [11] is similar to our methods as it determines characteristic keyphrases for each person entity and compares them to the context of a person mention. Another comparable approach [2] makes use of hyperlinks from Wikipedia and mention context using a dictionary, a graph, and textual context. The dictionary maps surface forms to a Wikipedia articles. This is similar to our approach but instead of using anchor texts, which need to be extracted from different Wikipedia pages, we make only use of Wikidata labels and alternative names. The graph in [2] is built from the Wikipedia link structure. In our approach, however, we calculate similarities between two person in the network by the number and distance of their wikilink co-occurrences within Wikipedia articles. Personalized PageRank and random walks are used in [2] to assign a Wikipedia article to a person mention. Our approach, in contrast, only relies on a social network that is constructed from co-occurrences of persons, leading to more efficient neighborhood computations and in particular better disambiguation results.

3. RESOURCES

The decision to use Wikimedia sites as knowledge backbone is based on several reasons: (1) the knowledge base is very large and not targeted towards a specific domain, (2) it deals with persons and communities that are mostly well-known, (3) the different projects can easily be combined, and (4) a variety of information for persons is available.

3.1 Wikidata

Wikidata is a free, collaboratively edited, multilingual database launched by the Wikimedia Foundation in 2012 [21]. As of January 26, 2015, Wikidata includes more than 16.8M items, which represent real life topics, concepts, and subjects. Each item is described by a unique identifier, a label, a description and statements that characterize the item. We extracted about 2.6M person entries that are classified as “instance of human” from Wikidata. Additional information, such as gender, date of birth/death, occupation, country of citizenship or site links to Wikipedia is provided. Several person names (variants, alternatives) can be associated with a Wikidata item. The *label* is the most commonly used name and in addition a list of alternative names (*aliases*) is available in different languages.

3.2 Wikipedia Social Network

We built a Wikipedia Social Network (WSN) [7] combining wikilinks (WL) in the English Wikipedia with person information in Wikidata. The text of about 5.3M content pages from the English Wikipedia[1] was cleaned from mark-up and split into sentences. WLs are links between Wikipedia pages. They are enclosed in double square brackets as in $[[linkTarget|coveredText]]$, where *coveredText* is optional. To identify WLs referring to persons, we use link information in Wikidata and category information in Wikipedia. If the *linkTarget* is equal to the English Wikipedia sitelink of a Wikidata person item, its id is assigned to the WL. The English Wikipedia contains about 76.8M WLs of which 10.4M refer to $842,484$ different persons. To find even more person mentions, each page is searched for all *linkTarget*s and *coveredText*s of its WLs, resulting in additional 2.6 M references to $273,166$ persons.

3.2.1 Network Construction

The WSN was built using co-occurrences of persons on Wikipedia pages (for full details see [7]). A bipartite graph of persons and documents is projected onto the set of persons to obtain a network of persons. In the resulting multigraph, each node represents a person and each edge a co-occurrence of the two connected persons. For each edge, a weight is calculated using a decaying distance measure which takes the number of sentences between the mentions into account. The multiple edges between nodes are aggregated using a cosine similarity of adjacency vectors of nodes in the weighted node-edge incidence matrix. This corresponds to a weighted cosine similarity of neighborhoods for the two incident nodes. The resulting person network[2] contains over 67M edges that connect $799,181$ different persons.

4. DISAMBIGUATION MODEL

In our approach, person mentions are disambiguated by comparing their entity candidates to uniquely identified persons in the same document, the so-called *seed persons*. We aim to infer the correct entity for a person mention by maximizing the relationship to the known neighborhood in the document. The overall approach is illustrated in Figure 1 and discussed in the following.

4.1 Extracting Person Mentions From Text

The first step is to detect person mentions (occurrences of names in a text document). In our approach the Stanford Named Entity Recognizer [6] is used to identify person mentions. Instead of clustering all person mentions referring to the same entity across several documents, our aim is to map these mentions to a reference list of uniquely defined person entities. This mapping makes is possible to include external knowledge like date of birth/death, occupation or affiliation.

[1] The English Wikipedia dump from Jan. 15, 2015 is used.

[2] The network is available for download from: `http://dbs.ifi.uni-heidelberg.de/index.php?id=data#WSN`

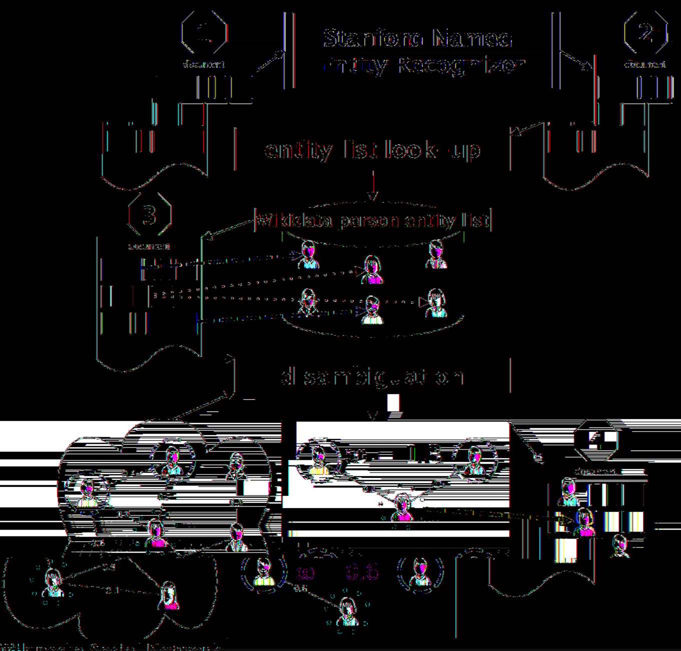

Figure 1: Overview of SoCCN4: The four stages of our disambiguation approach for person mentions.

4.2 Entity List Look-up

Once the person mentions in a document have been identified (stage 2 in Figure 1), the next step is to find possible matches in an entity list, which can be extracted from several sources. We exclusively use a list of Wikidata person items. We assume that every person mention refers to an entity in our database, since the knowledge base contains hundreds of thousands of person entities.

For the look-up (stage 3 in Figure 1), the surface form is compared to the names in the entity list using different string matching algorithms (for an overview, see [4]). For mentions that consist of more than one word, first exact string matching on the Wikidata label is used. If no database entry matches the mention, the search is extended to the alternative names. If again no possible candidate can be found the standard text search function in MongoDB is used[3], which ignores language specific stop words, matches on the complete stemmed word and returns a text score for each match. In our approach we consider only candidates with a text score > 0.7. For mentions containing only one word, the MongoDB text search is used on the Wikidata label only.

There are three possible outcomes of this list look-up: (1) no match was found, (2) exactly one match was found (unambiguous person mention), and (3) more than one match was found (ambiguous person mention). In the first case the person mention cannot be linked to an entity (the grey instances in Figure 1). In the second case (the green instances in Figure 1), the person mention is uniquely identified with one entity id. In the third case, more than one match (in the following *entity candidate*) has been found (the blue instance in Figure 1). This list of candidates C_m consists of those entities from the entity list that match the given person mention m.

4.3 Social Network Disambiguation

Given a list C_m of candidates for a person mention m in a document d, the task now is to determine the best match $c \in C_m$. For this, we employ a measure of the relationship(s) between each candidate $c \in C_m$ and other persons in the document. Key to this approach is that the "other persons" in the document are the unambiguous person mentions obtained during stage 2 above. We call them henceforth the *known neighborhood*, denoted N, which consists of so-called *seed persons sp*. Note that this neighborhood (and thus all seed persons) is independent of m as it solely consists of unambiguous person mentions in d. The entity candidate $c \in C_m$ for a given person mention m that is ranked highest is chosen as *identified entity* for the person mention m. To determine which candidate of a candidate list best fits the known neighborhood N, we use social network matching.

The WSN is a network of person entities extracted from the English Wikipedia. Each connection between two co-occurring persons is weighted by the (cosine) similarity in terms of number of co-occurrences and distances in their co-occurrences. That is, the connections of persons that co-occur frequently and near to each other on a Wikipedia page are assigned a higher weight than persons that co-occur rarely and/or far apart from each other on pages. The weight of the connection between two persons thus can be viewed as the strength of their relationship.

For each pair of persons (p_x, p_y), the strength of their relationship strength r is obtained from the edge weight in the WSN as a value between 0 and 1. If two persons are not connected in the WSN, they are assigned a relationship strength of 0; if $p_x = p_y$, then r has the value 1. We define a weight function $\omega : (c, N) \to \mathbb{R}$ for the neighborhood relation between a candidate $c \in C_m$ for a person mention m and a neighborhood N of seed persons sp in the document d. It is the sum of the relationship strengths r between the candidate c and every seed person sp of neighborhood N:

$$\omega(c, N) = \sum_{sp \in N} r(c_m, sp) \tag{1}$$

The entity candidate $c \in C_m$ with the maximum value of $\omega(c, N)$ is selected as the best fitting person entity and is the identified entity for the given person mention (stage 4 in Figure 1). Each person mention thus can be uniquely mapped to exactly one entity in the entity list.

What effect does the size of the candidate list and the neighborhood N in a document have on the quality and correctness of the mapping? We will address these questions, among others, in the following section.

5. EXPERIMENTS

For evaluating the quality of person name disambiguation using the WSN, we use wikilinks referring to persons (PWLs). Each of the 10M PWLs is linked to a unique Wikidata id (see Section 3.2). The Stanford NE recognizer finds about 41M person mentions in the English Wikipedia. The intersection of these two collections consists of more than 8M person mentions from which we use a subset of about 1.5M person mentions in $158,521$ Wikipedia pages, referring to $308,875$ different person entities as *ground truth*. Since every person mention in this subset was found by Stanford NER and is a wikilink, the unique entity to which this person mention refers is clearly known. Therefore, we can compare

[3] For details see the MongoDB 3.0 Manual at `http://docs.mongodb.org/manual/reference/operator/query/text/`

	SocNNEL		base$_1$		base$_2$	
	P	A	P	A	P	A
all mentions	96.4	94.1	89.0	88.9	78.5	78.4
explicit	98.7		98.7		98.7	
ambiguous	89.8	82.4	63.0		24.2	
multi-word mention	97.5	95.8	91.6	91.6	83.3	83.3
explicit	98.7		98.7		98.7	
ambiguous	92.7	85.6	66.5		29.2	
one-word mention	82.3	73.3	56.6	5.4	17.8	17.5
explicit	95.1		95.1		95.1	
ambiguous	80.5	72.3	51.7		8.0	

Table 1: Accuracy (A) and Precision (P) in % on a subset of the wikilink collection.

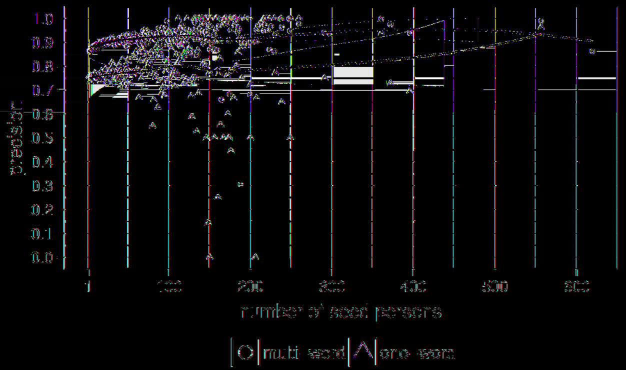

Figure 2: Precision in relation to the number of seed persons/person mention; the lines are the statistical regression lines with confidence intervals of 95%.

the entity that is assigned to a person mention by our disambiguation model to the correct entity.

The next step is to find possible candidates for each person mention. For only 0.17% of the person mentions, we do not find any entity candidate in the entity list.

The known neighborhood is then extracted for each document by identifying the seed persons (stage 3 in Figure 1). The relationship strength between each candidate for an ambiguous person mention and the known neighborhood is calculated. The candidate with the maximum relation strength is selected (stage 4 in Figure 1). 97.6% of all mentions were disambiguated and labeled with an entity id.

To estimate the quality of our person name disambiguation model the well-known IR measures micro precision (P) (aggregated over all mentions across all texts) and accuracy (A) are used following the definition in [11, 10]:

$$P = \frac{|correct|}{|found|}, \qquad A = \frac{|correct|}{|all|}$$

We achieve an overall accuracy of 94.12% with a precision of 96.43% (see Table 1). We compare our Social Network Named Entity Linking approach (SocNNEL) to two baselines. For base$_1$, the entity candidate with the lowest Wikidata id within the set of entity candidates was chosen. In Wikidata, well known and popular persons tend to have a lower id, since the ids are given consecutively by order of database entry. For *base$_2$*, the entity was randomly selected from the set of candidates. We took an average of 10 random selections to calculate P and A. In the following, we discuss the results in detail. Since the known neighborhood is essential for our approach, we will start with discussing the results for unambiguous person mentions – the seed persons and then continue with the ambiguous person mentions.

5.1 Known Neighborhood and Seed Persons

For our approach, the known neighborhood of a person mention plays a key role. If the seed persons are not correctly identified or only a few or none are found, the disambiguation model will struggle.

72.9% of the person mentions in the evaluation set are unambiguous and are used as seed persons. Only 1.3% of the seed persons are assigned to the wrong entity. For example, the correct entity for the surface form "John Thomas Graves" is "John T. Graves" the Irish mathematician, but instead it was assigned to "John Thomas Graves" the Confederate Army soldier. Similarly "Leo VI" was mapped to the pope "Leo VI" and not to the correct entity "Leo VI the Wise", a Byzantine Emperor. We tried to predict putative unambiguous surface forms that might be assigned to the wrong entity to apply different string matching algorithms. Our guess is that roman literals or initials might be the problem. But only 18% of the incorrectly assigned person mentions include a roman numeral and 3.6% an initial. Also, 85.5% of the seed persons containing roman numerals and 93.2% containing an initial were assigned to the correct entity. It is not advisable to use other (fuzzy) string matching algorithm for all surface forms, since this might lead to more ambiguous mentions, which decreases the number of seed persons. Since the error rate of 1,3% is very low, we consider this shortfall as acceptable.

5.2 Ambiguous Person Mentions

We are able to establish solid known neighborhoods as a basis for disambiguation. In this section, we will look at the quality of our disambiguation model. We consider the remaining 404,622 person mentions that can refer to several person entities. The overall precision for ambiguous person mentions is 89.8% with an accuracy of 82.41%.

For 6.8% of the ambiguous person mentions, the correct entity was not in the candidate list created during *entity list lookup*. Thus, it is not possible to assign the correct id. A solution for this problem might be to use other string matching algorithms or to always include the alternative names. But it is difficult to anticipate whether the correct entity is not listed in the set of candidates. If the alternative names are included in the standard look up or different string matching algorithms are used, the number of candidates will most likely increase while the number of seed person will decrease. An approach worth considering is to use a threshold. If the neighborhood relation ω of the best candidate is below a certain threshold, additional resources should be considered. For 4.8% of the ambiguous person mentions, there is no known neighborhood available and thus it is not possible to find the best fitting candidate.

If the mentions without a neighborhood or without the correct entity within the list of candidates are discarded, the adjusted accuracy is 92.6% and the precision is 92.8%. In other words, if the entity look up gives us a list of entity candidates where the correct entity is included and we find at least one seed person, then our model selects for more than 9 out of 10 mentions the correct entity. This number is quite remarkable considering that the WSN is only based on person co-occurrences within Wikipedia pages. No knowledge of the mention's context is needed (apart from other persons), no context similarity between the entity and the

	SocNEL		base1		base2	
	A	P	A	P	A	P
all m	84.4	94	71.2	[illegible]	62	61.9
truth possible	86.2	95.7	75.8	75.8	63.2	63.2
ambiguous m	65.8	87.6	43.1	43.1	14.5	14.5
truth possible	66	89.5	[illegible]	[illegible]	[illegible]	[illegible]

Table 2: Accuracy (A) and Precision (P) in % on [illegible]

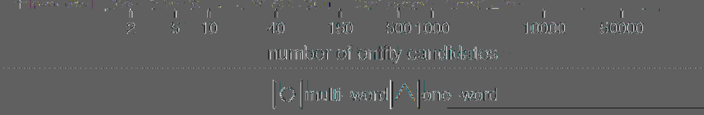

Figure 3: Precision in relation to the number of candidates per person mention. The lines are the statistical regression lines with confidence intervals of 95%.

mention has to be calculated, and no other external knowledge like biographical information is used.

In Section 4.2 we described two different methods for the entity list look-up. The person mentions are divided into the subsets multi word mentions (92,6% of all mentions) and one word mentions. The accuracy for the ambiguous multi word mentions is 85.6% with a precision of 92.7% (adjusted A= 95.2%, P= 95.3%) and for the ambiguous one word mentions it is A= 72.3%, while P= 80.5% (adjusted A= 84.0%, P= 84.2%). In the following, we analyze the impact of different parameters on these subsets: the number of seed persons and the number of candidates per mention.

5.2.1 Number of seed persons

The seed persons are very important for our approach. The question arising is whether the number of seed persons per document influences the disambiguation performance. The precision in relation to the number of seeds is shown in Figure 2. The regression lines indicate that the number of seeds does not have a huge impact on precision. The precision for one word mentions, which tends to be more influenced by the size of the neighborhood, is always lower than the precision of multi word mentions.

5.2.2 Number of candidates

Does the number of candidates per mention influence the precision? The assumption is that with a growing number of candidates the probability of selecting the correct entity decreases. For multi word mentions up to 64,980 and for one word mentions up to [illegible] candidates are found. The number of candidates has an impact on the precision. Figure 3 shows that for one word mentions, the precision drops considerably if more than 1000 candidates are found. This is not only due to the fact that with a higher number of possible entities the probability of selecting the correct one [illegible] multi-word mentions 10 candidates. The challenge here is to minimize the number of candidates while maintaining the ability to include the correct entity in the list of candidates. Thus the number of candidates needs to be filtered sensibly. One approach is to use external information about the candidates and the neighborhood, e.g., to select only those candidates that lived during the same time period as the seed persons.

To summarize, the social network approach to disambiguating person mentions provides reliable results in terms of precision and accuracy. The known neighborhood plays an important role for our method, but the actual number of seeds person does not greatly affect the quality. Thus even ambiguous mentions with a small known neighborhood can be correctly disambiguated with high precision. On the other hand the number of entity candidates per mention influences the precision, with less candidates resulting usually in higher precision. The experiments on the Wikipedia dataset gave us valuable insights into the functionality, power and limitations of our approach. In the next section, we will test our method on a standard dataset for NED.

6. EVALUATION

To compare our approach to other methods we evaluate it on a standard data set, namely the AIDA-CoNLL Dataset [illegible]. The test b set was used for evaluation. For the original set 4,485 mentions are annotated with a Yago2 entity and a Wikipedia link. We used the Wikipedia link information to map the mentions to Wikidata ids. Since in our current approach only person mentions are disambiguated, a subset of the *test b* set is used, which consists of 984 person mentions.

For 830 of the 984 mentions our approach returns the correct entity, which leads to an accuracy of 84.4% with a precision of 94%. For 101 of the mentions, the proposed method does not find a matching entity. For one of these mentions, no candidate is found in the entity list, and for 53 mentions there is no known neighborhood available. Although the WSN consists of more than 57 M edges, for 47 of these mentions, there is no connection between any possible candidate and any seed person.

The performance of our method depends on two crucial steps: (1) finding candidates in the database, and (2) selecting the correct entity from this list of candidates. [illegible] of the mentions, the correct entity was not on the candidate list. Therefore, the second step cannot select the correct entity. Excluding these mentions results in an overall adjusted accuracy of 86.2% while P = 95.7% For the 422 ambiguous mentions the accuracy is 65.8% with a precision of 87.6%. [illegible] pare the results. The AIDA dataset includes different types of named entities (person, locations, organizations) whereas our method focuses on person mentions only. Unfortunately, we were not able to find any information on the distribution of the different types within the dataset. In the evaluation of the other approaches no distinction is made between the dif-

	A	Micro P	MAP
SocNNEL	84.4	94	95.2
Hoffart et al. 2011 [11]	82.5	81.9	89.1
Houlsby & Ciaramita 2014 [12]	84.9		
Moro et al. 2014 [19]	82.1		
Barrena et al. 2015 [2]	83.6		

Table 3: Accuracy (A) in % on AIDA CoNLL-YAGO Dataset test-b.

ferent types of entities. Secondly, it is not clear if the results of the mentioned methods are calculated over all mentions or only the ambiguous mentions. Thirdly, in contrast to the other approaches, our method does not always assign an entity.

In conclusion it can be said that our system achieves state-of-the-art performance with regard to accuracy, and it outperforms the AIDA system [11] in micro precision and MAP. This is remarkable since our system is just based on co-occurrences of person names in Wikipedia pages. The coherence among entities and the context similarity of mention and entity as in [11] are not taken into account, instead we rely only on the help of the neighborhood.

7. CONCLUSION AND ONGOING WORK

In this paper we presented an approach for named entity linking. We introduced a disambiguation model that employs information about the neighborhood of person names in a document and a large knowledge base: the Wikipedia Social Network. We established a weight for the relationship between a person and its neighborhood, which takes the relationship strength of two persons based on their co-occurrences in Wikipedia into account. With a large scale evaluation on more than 1.5 M person names we showed that our approach yields an overall precision of over 96% for person name linking. The effectiveness of our disambiguation model is proven by the precision of over 89% for selecting the correct person entity for an ambiguous person mention. We showed that the Wikipedia Social Network is a valuable resource for named entity linking and that our approach is well-positioned when compared to other state-of-the-art methods. On the standard dataset for NED we received an accuracy of over 84% and a precision of 94%, which is considerably higher than in comparable approaches that use extensive knowledge on context and coherence. Our method using the Wikipedia Social Network based on co-occurrences is reliable and simple. It can be applied to different document types and is easily adoptable to other languages. We are working on building Wikipedia (Social) Networks for different languages and different named entities, for example, for place names [8]. We are currently refining our method in order to cover more person mentions and to become more independent from seed persons. We are also looking into intelligently limiting the number of candidates for person mentions by using external knowledge.

8. REFERENCES

[1] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the Vector Space Model. In *COLING*. ACL, 1998.

[2] A. Barrena, A. Soroa, and E. Agirre. Combining Mention Context and Hyperlinks from Wikipedia for Named Entity Disambiguation. In **SEM*. ACL, 2015.

[3] R. Bunescu and M. P. Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *EACL*, 2006.

[4] P. Christen. A Comparison of Personal Name Matching: Techniques and Practical Issues. In *ICDMW*. IEEE, 2006.

[5] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *EMNLP-CoNLL*. ACL, 2007.

[6] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL*. ACL, 2005.

[7] J. Geiß, A. Spitz, and M. Gertz. Beyond Friendships and Followers: The Wikipedia Social Network. In *ASONAM*. IEEE, 2015.

[8] J. Geiß, A. Spitz, J. Strötgen, and M. Gertz. The Wikipedia Location Network - Overcoming Borders and Oceans. In *GIR*, 2015.

[9] X. Han and J. Zhao. Named entity disambiguation by leveraging wikipedia semantic knowledge. In *CIKM*. ACM, 2009.

[10] J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *WWW*. ACM Press, 2014.

[11] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *EMNLP*, 2011.

[12] N. Houlsby and M. Ciaramita. A Scalable Gibbs Sampler for Probabilistic Entity Linking. In *Advances n Informat on Retr eval (ECIR)*, 2014.

[13] M. Ikeda, S. Ono, I. Sato, M. Yoshida, and H. Nakagawa. Person name disambiguation on the web by two-stage clustering. In *WePS, 18th WWW Conference*, 2009.

[14] D. V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra. Towards breaking the quality curse.: a web-querying approach to web people search. In *SIGIR*, 2008.

[15] B. Malin, E. Airoldi, and K. M. Carley. A network analysis model for disambiguation of names in lists. *Comput. Math. Organ. Theory*, 11(2), 2005.

[16] G. S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *CoNLL*, 2003.

[17] D. N. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM*, 2008.

[18] E. Minkov, W. W. Cohen, and A. Y. Ng. Contextual search and name disambiguation in email using graphs. In *SIGIR*, 2006.

[19] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *TACL*, 2014.

[20] J. Tang, Q. Lu, T. Wang, J. Wang, and W. Li. A bipartite graph based social network splicing method for person name disambiguation. In *SIGIR*, 2011.

[21] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Commun cat ons of the ACM*, 57(10), 2014.

[22] M. Yoshida, M. Ikeda, S. Ono, I. Sato, and H. Nakagawa. Person name disambiguation by bootstrapping. In *SIGIR*, 2010.