Extending AIDA framework by incorporating coreference resolution on detected mentions and pruning based on popularity of an entity

Samaikya Akarapu
Department of Computer Science
and Engineering
Indian Institute of Technology (BHU)
Varanasi, India 221005
samaikyaakarapu@gmail.com

C Ravindranath Chowdary
Department of Computer Science
and Engineering
Indian Institute of Technology (BHU)
Varanasi, India 221005
rchowdary.cse@iitbhu.ac.in

Abstract

Named Entity Disambiguation (NED) is gaining popularity due to its applications in the field of information extraction. Entity linking or Named Entity Disambiguation is the task of discovering entities such as persons, locations, organizations, etc. and is challenging due to the high ambiguity of entity names in natural language text. In this paper, we propose a modification to the existing state of the art for NED, Accurate Online Disambiguation of Entities (AIDA) framework. As a mention's name in a text can appear many times in shorter forms, we propose to use coreference resolution on the detected mentions. Entity mentions within the document are clustered to their longer form. We use the popularity of candidate entities to prune them and based on the similarity measure of AIDA the entity for a mention is chosen. The mentions are broadly classified into four categories person, location, organization and miscellaneous and the effect of coreference and pruning were analyzed on each category.

1 Introduction

One of the unsolved problems in computer science is understanding and producing natural language by machines. The goal of fully understanding is out of reach but there have been significant advances recently. Systems are able to understand words or phrases of text by explicitly representing their meaning. Once the meanings of individual words are known, next is to find the relation among them.

1.1 Named Entity Disambiguation

A real word object that is designated by a proper name that identifies itself from other objects having similar attributes is called as named entity. Names can be rigid or non-rigid. Rigid names refer to one and only one thing like "*Narendra Modi*". Non-rigid names refer to different objects like "*Home Minister*" (Home Minister of India or Srilanka). In general, we can say proper names are rigid and common names are non-rigid. Articles on the web consist of names of persons, locations, organizations, events etc. The same name can have a different meaning. For example, consider the following sentence:

Example 1.1 *" Michael is the father of two relational database systems, Ingres and Postgres developed at Berkeley. Page and Brin did research at Stanford."*

Here "*Michael*" refers to the person *Michael Stonebraker* who is a computer scientist and not the singer *Michael Jackson*, "*Berkeley*" and "*Standford*" refer to the universities- *University of California, Berkeley* and *Standford University* and not to the places *Berkeley* and *Standford*, "*Page*" refers to *Larry Page* the founder of Google and not *Jimmy Page* who is a guitarist. Looking at the sentence, humans barely notice the ambiguity as they subconsciously resolve it. The ability to understand single words was made possible by associating phrases and words with their senses. The WordNet (Fellbaum, 1998) contains a collection of senses for nouns, adjectives and verbs and word sense disambiguation (Navigli, 2009) has benefited from it. Mapping these mention names to the actual entities is referred to as Entity Linking or Named Entity Disambiguation.

1.2 Named Entity Recognition

Before applying NED, the first step would be to recognize a word or multiple word phrases that could possibly represent a real word entity. For the last two decades, entity recognition has received a lot of attention. The task of finding and categorizing elements of text into different classes

D S Sharma, R Sangal and A K Singh. Proc. of the 13th Intl. Conference on Natural Language Processing, pages 36–45,
Varanasi, India. December 2016.

such as names of persons, locations, organizations, quantities, expressions of times, percentages, monetary values, etc. is termed as Named Entity Recognition or entity identification, in short as NER. Most NER methods use machine learning to label the input texts. The data for the training is mostly obtained from MUC (Message Understanding Conference) (Grishman and Sundheim, 1996), where NER was first introduced, and CoNLL (Computational Natural Language Learning) (Tjong Kim Sang and De Meulder, 2003). The most widely used system for NER is the Standford NER (Finkel et al., 2005) that uses the conditional random fields.

1.3 Representation

The NED maps ambiguous names to its canonical entities. It assumes that the names or phrases that could potentially represent a real world entity are discovered by using a NER. These names are called as mentions. M are the set of mentions which are given as input. KB is the knowledge base that is used as the reference list of entities. E is the set of entities. If Wikipedia is taken as a knowledge base, each page of the Wikipedia is an entity. $m \in M$, $D \subset (N \times E)$ is the dictionary that contains the pairs of (n, e) where n is a name $\in N$ and $e \in E$. N is the set of all names of each e. Suppose if the entity *Michael Jackson* is considered *MJ, Michael Joseph Jackson, King of Pop* etc. would be the set of names for this entity. $CE(m)$ are the candidate entities for a mention m. To find $CE(m)$, m is matched against names in N. The goal of NED is to map m to an entity in $CE(m)$. If the entity is not in the knowledge base it is mapped to *NULL*, i.e. it is not registered. If $CE(m)$ is empty by default m is mapped to *NULL*.

2 Related Work

NED requires a knowledge base to map the mention names to the corresponding entities registered in a specific knowledge base. One of the popular choices of a knowledge base is Wikipedia, where each page is considered as an entity. Bunescu and Pasca (2006) were the first to use Wikipedia to link entities. The basis of disambiguation is to compare context of mention and candidate entities. Milne and Witten (2008), Kulkarni et al. (2009) also considered the semantic relations between the candidate entities for disambiguation.

Pershina et al. (2015) represented NED as a graph model and disambiguated based on PersonalizedPageRank(PPR). The local similarity score includes the similarity between Wikipedia title, mention and category type. The global similarity is measured based on the link counts of Freebase and Wikipedia. Either of these measures is assumed as the initial similarity score. The coherence of entity is obtained as a pairwise relation of PPR scores with entities of other mentions. The final score of entity is a combination of coherence and initial similarity score weighted with PPR average. The entity with the highest score is selected and evaluation was done on CoNLL 2003 dataset used by Hoffart et al. (2011).

Luo et al. (2015) jointly recognize and disambiguate entities by identifying the dependency between the tasks (JERL). It defines three feature sets on a segment assignment of a word sequence: NER features (various unigram and bigram features, dictionaries, WordNet clusters, etc.), linking features (entity priors, context scores), mutual dependency (type-category correlation) and is modeled as Semi-CRF. Evaluation was done on CoNLL 2003 dataset used by Hoffart et al. (2011).

Usbeck et al. (2014) (AGDISTIS) finds an assignment that maximizes similarity with the named entities and coherence with the knowledge base. The candidate entities are found using trigram similarity and belong to categories of person, place, and organization. With candidate entities as initial vertices in a graph, it is expanded by DFS with a certain depth. The edge between vertices is present if they form an RDF triplet. They use HITS algorithm (Kleinberg, 1999) to find the authoritative candidates, sort them and assign them. They evaluated on eight different datasets: Reuters-21578 (Röder et al., 2014), news.de (Röder et al., 2014), RSS 500 (Röder et al., 2014), AIDA-YAGO2 (Hoffart et al., 2011), AIDA/CoNLL-TestB (Hoffart et al., 2011), AQUAINT (Hoffart et al., 2011), IITB (Kulkarni et al., 2009), and MSNBC (Cucerzan, 2007).

Moro et al. (2014) address Entity Linking and Word Sense Disambiguation. They consider the semantic network Babelnet (Navigli and Ponzetto, 2012), where each concept and named entity is a vertex and relations between them are edges. They perform Random Walk with Restart (Tong et al., 2006) to reweigh the edges and to obtain a semantic signature. In the input document, text frag-

ments which contain noun and substrings of entities in Babelnet are considered as candidate entities. Edges are added between candidate entities based on the previously computed semantic signature. A dense subgraph is found and the candidates are selected based on the score obtained from incident edges. They evaluated on two datasets for NED: KORE50 (Hoffart et al., 2012) and CoNLL 2003 dataset used by Hoffart et al. (2011).

Almost all the methods use the similar features to find similarity between the context of mentions and their candidate entities but differ in disambiguation method. The methods can be broadly divided into two types: local method and global method. While disambiguating, the local method only considers mention and its candidate entities but the global method also consider relations among the entities. Thus, the complexity of the global method is high. The systems assume the annotations are correct if they strictly match the ground truth. The difficult part for the systems would be disambiguating entities that don't belong to Wikipedia as the features of these entities are absent. All the above methods use different data sets for evaluation. Pershina et al. (2015), Luo et al. (2015) did not consider the assignment of null entities. Usbeck et al. (2014) used DBpedia as knowledgebase but with Yago2, AIDA performed well.

3 AIDA

AIDA (Hoffart et al., 2011) is a framework developed by the *Databases and Information Systems Group at the Max Planck Institute for Informatics* for named entity recognition and disambiguation. The framework presents both local and global disambiguation methods for NED. In the local disambiguation technique, the disambiguation is done based on prior probability and context similarity with a prior test. The prior test decides whether prior probability has to be considered or not. In the global disambiguation technique, the NED is presented as a graph problem with mentions and entities as nodes and weighted edges between them. An edge is present between mention and its candidate entities. An edge between e_1 and e_2 is present if they are candidate entities of different mentions and have a link to their pages. The edge weight between a mention and an entity is the similarity between the context of mention and context of the entity. The edge weight between entities is proportional to their coherence. Not all mentions are disambiguated by the global method. A coherence test decides whether the disambiguation should be done locally or globally. The goal is to find a subgraph with each mention having only one edge with an entity thus disambiguating collectively. Collective disambiguation was proposed by Kulkarni at el. (2009). The AIDA was evaluated on 1393 articles of CoNLL 2003 dataset and mentions were recognized using Standford NER (Finkel et al., 2005) tagger. The following features of AIDA are used later in the experiments:

Prior Probability: Popularity, in general, gives an estimate of what a mention could be referring to. The measure is obtained based on Wikipedia link anchors. The probability distribution of candidate entities is estimated as the number of times the entity referred with that mention as the anchor text in Wikipedia.

KeyPhrase-based Similarity: The important measure for mapping is the similarity between the context of mention and entity. All the tokens in the document are considered as the context of the mention. The keyphrases extracted from Wikipedia link anchor texts, category names, citation titles, external references of the entity and the entities linking to it are considered as the context of the entity. These are the set of keyphrases of entity $KP(e)$. The Mutual Information **MI** between an entity e and a word w occurring in a keyphrase is calculated as (Hoffart et al., 2011):

$$MI(e,w) = \frac{\#elements\ in\ KP(e)\ in\ which\ w\ occurs}{N} \quad (1)$$

where N is the total number of entities.

Each keyphrase q in $KP(e)$ is associated with a score by obtaining a *cover(q)*- the smallest window in the text such that maximum number of words in q occur in the window. The score of a phrase q is given by (Taneva et al., 2011):

$$score(e,q) = z_q \left(\frac{\sum_{w \in \{cover(q) \cap q\}} MI(e,w)}{\sum_{x \in q} MI(e,x)} \right)^2 \quad (2)$$

where $z_q = \frac{\#\ matching\ words\ in\ cover(q)\ and\ q}{length\ of\ cover(q)}$.

The similarity between mention and a candidate entity is given as the sum of scores of all keyphrases of entity (Hoffart et al., 2011):

$$sim(e) = \sum_{q \in KP(e)} score(e,q) \quad (3)$$

4 Coreference and Pruning

Coreference resolution is defined as finding all expressions that refer to the same entity in a text. In a text, it can happen that one of the names of an entity is long and later the same names of the entity are referred with short forms. Our concern is coreference resolution on the mentions detected in the text. For example,

> "*Sir Jagadish Chandra Bose* ↑ *is one of the fathers of radio science.* *Bose* *was the first to use semiconductor junctions to detect radio signals.*"

The idea is to map the shorter forms to the longer forms. Longer forms are more explicit and can have few candidate entities or just one as compared to the shorter forms.

We use the Standford NER tagger to obtain the tokens and their labeling. A mention is a span of token/tokens. AIDA also uses Standford NER for detecting the mentions. A mention phrase is found if the span of token/tokens has the same label. Thus a mention is labeled accordingly into one of the four categories person, location, organization and miscellaneous. While coreferencing, the labeling information is used. The shorter forms are mapped to the longer form of a mention if the label of both the forms are same and the shorter form occurs in the longer form. Consider the following example:

> "*Ram Prasad, designer of .... Shiva Prasad works as a doctor at AIMS.... Dr. Prasad ....*"

here *"Dr. Prasad"* refers to *"Shiva Prasad"*. The way we match would map *"Prasad"* to *"Ram Prasad"*, since the mapping is based on text matching. To map *"Prasad"* to *"Shiva Prasad"* the context *"Dr."* should be considered. But usually in a text, if people with same family name appear they would be referred using their first name, so the matching is kept to a simple string matching. The condition that the labeling should be same is imposed to ensure that the short name occurring in the long name but belonging to different entities are not clustered. For example,

> "*Universtiy of Delhi is a central collegiate university, located in Delhi.*"

"University of Delhi" is an organization and *"Delhi"* is a place. *"Delhi"* occurs in *"University of Delhi"*. If the condition is not imposed both names of different entities would be clustered which is incorrect. The experiments were done both imposing and not imposing the condition.

At the next stage, the mentions are queried for the candidate entities. For all the candidate entities of a mention, prior probability $prior(e)$and keyphrase-based similarity (not the mention-entity similarity which may include prior probability) $sim(e)$ is obtained using AIDA. The mentions which have only one candidate entity are mapped to it directly and which have no candidate entities are mapped to null. We calculate the global average of the prior probability of the candidate entities of all mentions, local average of the prior probability of the candidate entities of each mention.

$$global\_avg_M = \frac{\sum_{m \in M} \sum_{e \in CE(m)} prior(e)}{\sum_{m \in M} size(CE(m))} \quad (4)$$

$$local\_avg_m = \frac{\sum_{e \in CE(m)} prior(e)}{size(CE(m))} \quad (5)$$

where m is a mention, M is the mention set, e is an entity, $CE(m)$ is the set of candidate entities of m, $size(CE(m))$ is the number of candidate entities of m, $prior(e)$ is the popularity or prior probability of e.

The candidate entities whose prior probability is lower than either $global\_avg_M$ or $local\_avg_m$ are pruned. Among the candidate entities left, the one with the highest keyphrase-based similarity $sim(e)$ is associated to the mention. Pruning is done to remove those entities whose popularity is very low as compared to those entities whose popularity and similarity are reasonably high. Popularity captures a notion of commonness. The frequency of occurrence of entities varies for different categories. On an average, places tend to occur more frequently or more popular than persons. Experiments were done with and without pruning and its trend in each category is examined and decided whether it should be applied or not. Later coreference and pruning are combined.

4.1 Finding and labeling the mentions:

Standford NER takes text as input and gives tokens $t_1, t_2, ...t_n$ and their labels $t_1.label$, $t_2.label..t_n.label$ as output. A mention is a span of tokens whose labels are same and the mention type is the label type, i.e. $m = \{t_k...t_{k+l} \mid t_k.label = ... = t_{k+l}.label\}$ and $m.label = t_k.label$.

4.2 Mapping short forms on to longer forms:

A mention is a span of tokens. A mention $m_1 = t_i...t_{i+p}$ is mapped to a mention $m_2 = t_j....t_{j+q}$ if m_1 occurs in m_2, i.e. $t_i = t_k \wedge ... \wedge t_{i+p} = t_{k+p} \wedge j <= k <= j+q-p \ \wedge \ m_1.label = m_2.label$.

Algorithm 1 NED

Input: Text
Output: Mention Mappings
Find mentions and label them as discussed in Section 4.1;
Map short forms on to longer forms as discussed in Section 4.2;
for m **do**
  if $CE(m) == null$ **then**
    $result\_entity(m) \leftarrow null$;
  end if
  if $size(CE(m)) == 1$ **then**
    $result\_entity(m) \leftarrow e$;
  end if
  if $size(CE(m)) > 1$ **then**
    for $e \in CE(m)$ **do**
      if $prior(e) < min(local\_avg_m, global\_avg_M)$ **then**
        $CE(m) \leftarrow CE(m) - \{e\}$;
      end if
    end for
    $result\_entity(m) \leftarrow \{e_i \mid \underset{i}{argmax}\ sim(e_i)\}$;
  end if
end for

4.3 Datasets

The experiments were carried on two data sets. First one is the TIPSTER[1] data set from which 45 documents were randomly chosen. These documents were related to news. The second dataset is the IITB dataset by Kulkarni et al.(2009), out of which 50 documents were taken. The IITB documents were collected from online news sources and are not well formatted and sometimes had comments of online users. The CoNLL 2003 dataset used by the AIDA is copyright protected but the annotations are available. We have manually annotated all the documents, i.e. both 45 documents of Tipster dataset and 50 documents of IITB dataset. The properties of the dataset are given in Table 1.

[1] http://www.nist.gov/tac/data/data_desc.html#TIPSTER

	Tipster	IITB
Documents	45	50
Mentions retrieved	1681	1666
Relevant mentions	1661	1595
Average mentions per document	37	32
Mentions marked as null in ground truth	383	457
Mentions not null	1278	1138
Mentions whose query resulted null	207	319

Table 1: Dataset Properties

	null	not null	total	%
p	240	337	577	34.74
l	48	621	669	40.28
o	58	206	264	15.89
m	37	114	151	9.09
total	383	1278	1661	
%	23.06	76.94		

Table 2: Mentions whose entity mappings marked as NULL and not NULL for Tipster dataset

Only the relevant mentions retrieved by AIDA are considered. For example, "*.... Bermuda-based company ....*" where AIDA retrieves "*Bermuda-based*" as mention which is considered irrelevant. Similarly, "*....Cuban-Soviet friendship...*" is retrieved as "*Cuban-Soviet*" which is irrelevant.

The mention mappings can be of four types: A mention whose entity is not registered in the database and mapping gives NULL, a mention whose entity is not registered in the database and maps to some entity, a mention whose entity is registered in the database and maps to an incorrect entity and mention whose entity is registered in the database and maps to the correct entity. Precision is the fraction of mention entity mappings that match the ground truth assignments. Macro average precision is the average of precision of each document. Micro average precision is the fraction of mention entity mappings in all documents that match the ground truth assignments. Table 2 and Table 3 give the details of NULL entities in ground truth. The recall remains the same for the AIDA and for the experiments done as both use the same retrieval methods. The mentions were retrieved using the Stanford NER which classifies the men-

	null	not null	total	%
p	239	149	388	24.33
l	35	580	615	38.56
o	145	330	475	29.78
m	38	79	117	7.36
total	457	1138	1595	
%	28.65	71.35		

Table 3: Mentions whose entity mappings marked as NULL and not NULL for IITB dataset

	p	l	o	m	total
p	563	7	5	2	577
l	3	522	21	123	669
o	3	12	231	18	264
m	4	2	26	119	151

Table 4: Confusion Matrix for Tipster dataset

tions into four categories namely person (p), location (l), organization (o) and miscellaneous (m). The mentions were also annotated for their labels manually. Table 4, Table 5 gives the Confusion Matrix for both the datasets. The column is the actual label and the row is the labels predicted by NER.

4.4 Experiments

AIDA was run with four settings:

- *LocalDisambigautionSetting()*- uses prior and similarity with a prior test, described in Section 3;
- *LocalDisambiguationWithNullSettings()*- uses the above method but uses a threshold to find NULL entities;
- *CocktailDisambiguationSettings()*- uses the graph method, described in Section 3;
- *CocktailDisambiguationWithNullSettings()*- uses the above method but uses a threshold to find NULL entities;

	p	l	o	m	total
p	374	8	3	3	388
l	27	494	32	62	615
o	27	60	360	28	475
m	6	2	22	87	117

Table 5: Confusion Matrix for IITB dataset

The experiments were carried out with the following 8 methods:

- **Method 1 (AG):** AIDA graph Disambiguation.
- **Method 2 (AGN):** AIDA graph Disambiguation with NULL settings.
- **Method 3 (AL):** AIDA local Disambiguation.
- **Method 4 (ALN):** AIDA local Disambiguation with NULL settings.
- **Method 5 (NP):** No Pruning- The method is run based on Algorithm 1 except the lines 4, 13 to 18 are not performed.
- **Method 6 (WP):** With Pruning- The method is run based on Algorithm 1 except the line 4 is not performed.
- **Method 7 (CNCP):** Coreference without labeling condition and pruning- The method is run based on Algorithm 1 but in line 4 short forms are mapped to longer forms of mention without the condition that both the forms should have the same label and pruning is done for all mentions labeled as location, organization, misc and not done for mentions labeled as persons.
- **Method 8 (CCP):** Coreference with labeling condition and pruning- The method is run based on Algorithm 1 but in line 4 short forms are mapped to longer forms of mention with the condition that both the forms should have the same label and pruning is done for all mentions labeled as location, organization, misc and not done for mentions labeled as persons.

Table 6 and Table 7 show the number of correct mappings for each category by various methods on the two datasets used. Among the AIDA methods, on the Tipster dataset, the graph disambiguation performs well. When considered for individual categories, it performs well on person and organization, while local disambiguation performs well on location and misc. On the IITB dataset, the local disambiguation performs well. When considered for individual categories, graph disambiguation performs well on person and organization, local disambiguation performs well on location and

	AL	ALN	AG	AGN	NP	WP	CNCP	CCP
person null	44.58	52.92	44.58	61.25	44.58	44.58	**83.75**	82.92
location null	**70.83**	**70.83**	**70.83**	**70.83**	**70.83**	**70.83**	**70.83**	**70.83**
organization null	82.76	82.76	82.76	84.48	82.76	82.76	**86.21**	84.48
misc. null	37.84	37.84	37.84	**40.54**	37.84	37.84	**40.54**	**40.54**
person not null	84.57	82.20	**95.55**	92.58	85.46	81.01	93.77	94.66
location not null	88.89	88.89	84.70	84.70	82.14	**91.79**	87.28	90.34
organization not null	86.89	86.89	**88.83**	**88.83**	83.50	85.44	84.47	84.47
misc. not null	65.79	65.79	64.04	64.04	57.89	**69.30**	68.42	68.42

Table 6: Percentage of correct mappings for each category by various methods on Tipster dataset

	AL	ALN	AG	AGN	NP	WP	CNCP	CCP
person null	53.56	61.51	53.56	66.95	53.56	53.56	**73.64**	**73.64**
location null	54.29	54.29	54.29	54.29	54.29	54.29	**71.43**	**71.43**
organization null	70.34	70.34	70.34	70.34	70.34	70.34	**72.41**	71.72
misc. null	**71.05**	**71.05**	**71.05**	**71.05**	**71.05**	**71.05**	**71.05**	**71.05**
person not null	85.91	85.91	**94.63**	93.29	86.58	79.87	93.96	93.96
location not null	85.86	85.86	82.59	82.07	81.90	**87.59**	87.07	87.41
organization not null	79.70	79.70	**81.56**	75.168	78.48	74.24	73.03	73.03
misc. not null	**72.15**	**72.15**	56.96	56.96	62.03	68.35	68.35	68.35

Table 7: Percentage of correct mappings for each category by various methods on IITB dataset

misc. Thus, AIDA graph disambiguation performs well for person and organization while local disambiguation performs well for location and misc.

The method 5 (NP) maps the mentions with the highest similarity, without considering the prior probability. When the method 6- with pruning is compared with method 5- no pruning, for Tipster dataset there is an increase in accuracy for location, organization and misc and decrease for person. For IITB dataset there is an increase in accuracy for location and misc and decrease for person and organization. So for method 7 (CNCP) and method 8 (CCP) pruning was done for location, organization, misc. Table 8 and Table 9 shows the results for various methods for both the datasets. Comparing the results, Coreference helps increase the accuracy of mapping especially for person, pruning for location while AIDA performs well on organization, for misc pruning increased accuracy on Tipster dataset. Pruning decreases the accuracy for organization showing that some potential entity that could be mapped is removed. Coreferenece decreases accuracy for location on Tipster dataset while not much effective on IITB dataset because the shorter forms accuracy depends on the longer form it is mapped to. For misc every method is equally competitive. After the modification, the mentions whose longer forms are mapped as NULL are ensured that the shorter forms too are mapped as NULL but in the case of AIDA, the shorter forms were mapped to some other entity in Yago2. For mentions whose longer forms are mapped to the right entity, the shorter forms are mapped to the right entities by both the methods AIDA and CCP. In one of the documents,

> "*Naomi Foner, who wrote.... Her own experiences made Foner.....*"

AIDA just gives only *Naomi Foner Gyllenhaal* as candidate entity of mention "*Naomi Foner*", but the mention "*Foner*" doesn't contain *Naomi Foner Gyllenhaal* as one of its candidate entity. This might be because of some error in retrieval by AIDA. Coreferencing them ensured that mention "*Foner*" is mapped to the right entity. If the longer surface form is mapped to a wrong entity, then all the shorter forms too are mapped to the wrong entity. Thus, the accuracy depends on the mapping of longer forms.

> "*Nicholas Calas a poet and..... Calas...*"

The longer form "*Nicholas Calas*" has no candidate entities and shorter form "*Calas*" has been mapped to the right entity *Nicolas Calas*. Here

Method	person	location	organization	misc	true	Macro (%)	Micro (%)
AG	429	560	231	87	1307	77.22	78.69
AGN	459	560	**232**	88	1339	79.45	80.61
AL	392	586	227	89	1294	76.41	77.90
ALN	404	586	227	89	1306	77.19	78.63
NP	395	544	220	80	1239	73.84	74.59
WP	380	**604**	224	**93**	1301	77.27	78.33
CNCP	517	576	224	**93**	1410	85.41	84.89
CCP	**518**	595	223	**93**	**1429**	**86.23**	**86.03**

Table 8: Results of various methods on Tipster dataset

Method	person	location	organization	misc	true	Macro (%)	Micro (%)
AG	269	498	**371**	72	1210	79.72	75.86
AGN	299	495	350	72	1216	80.07	76.24
AL	256	508	365	**84**	1213	79.53	76.05
ALN	275	508	365	**84**	1232	80.69	77.34
NP	275	508	365	84	1232	80.69	77.34
WP	247	527	347	81	1202	79.21	75.36
CNCP	**316**	530	346	81	1273	82.52	79.81
CCP	**316**	**532**	345	81	**1274**	**83.03**	**79.87**

Table 9: Results of various methods on IITB dataset

coreferencing shorter form *"Calas"* to longer form maps it to NULL. Instead of *"Nicholas"* if it had been *"Nicolas"* it would had mapped to right entity. The mention was just misspelled.
The coherence graph algorithm of AIDA makes sense.

> *"....Maj. Gadi, commander of a battalion...."*

When only similarity is considered it maps to *"Gadi Brumer (Israeli footballer)"* but with coherence, it maps to *"Gadi Eizenkot (Chief of general staff of Israel Defence Forces)"*. Coherence too causes errors. All mentions with the same syntax are mapped to the same entity.

> *"...all Chinese in Tibet stop carrying... The demonstrators were carrying banners in Tibetian and Chinese..."*

Here the former *"Chinese"* means *Chinese people* and later means the *Chinese language*.

> *"...Jewish dietary laws....intones the Hebrew words.."*

both *"Jewish"* and *"Hebrew"* are mapped to *Hebrew language*.
"...The runners-up for the charisma title were San Francisco and Washington ...", here, *"San Francisco"* should be mapped to the teams *San Francisco 49ers* and *"Washington"* to *Washington Redskins* but are mapped to places. These entities occur at the top when sorted with respective to the similarity measure. If it is known that these mentions represent a team (organization), other candidate entities which are not team (organization) could be pruned by finding the yago types.

5 Conclusions

The proposed modifications to the AIDA improved the overall accuracy of entity mappings. The first modification is mapping short forms on to their long form. As long forms are more explicit, they are less ambiguous. It especially helps in identifying null entities, about an increase in 17.58% for Tipster dataset and 5.25% for IITB dataset. The second modification is pruning based on popularity. Experiment results show that applying pruning selectively on categories help in increase of accuracy of the system.

6 Acknowledgement

This project was supported by DST-SERB No. YSS/2015/000906.

References

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 9–16, Trento, Italy. Association for Computational Linguistics.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June. Association for Computational Linguistics.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 1*, COLING '96, pages 466–471, Copenhagen, Denmark. Association for Computational Linguistics.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Edinburgh, United Kingdom. Association for Computational Linguistics.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 545–554, Maui, Hawaii, USA. ACM.

Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 457–466, Paris, France. ACM.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal, September. Association for Computational Linguistics.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, Napa Valley, California, USA. ACM.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69, February.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized page rank for named entity disambiguation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 238–243, Denver, Colorado, May–June. Association for Computational Linguistics.

Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N3-a collection of datasets for named entity recognition and disambiguation in the nlp interchange format. *9th LREC*.

Bilyana Taneva, M Kacimi El Hassani, and Gerhard Weikum. 2011. Finding images of rare and ambiguous entities. *Technical Report*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Edmonton, Canada. Association for Computational Linguistics.

Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast random walk with restart and its applications. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 613–622, Washington, DC, USA. IEEE Computer Society.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer, and Andreas Both. 2014. Agdistis - graph-based disambiguation of named entities using linked data. In *Proceedings of the*

13th International Semantic Web Conference - Part I, ISWC '14, pages 457–471, New York, NY, USA. Springer-Verlag New York, Inc.