From Text to Knowledge: Semantic Entity Extraction using YAGO Ontology

Farhad Abedini, Fariborz Mahmoudi, and Amir Hossein Jadidinejad

***Abstract*—Semantic entities are the entities that their concepts are available in a knowledgebase. Here, a new system will be introduced to extract semantic entities from texts. For this aim a new disambiguation method is suggested to match each of ambiguous entity with one of semantic entities in the knowledgebase. The YAGO ontology is used in this method as state of the art of knowledgebase in this field. Since entities in YAGO are meaningful, so in this method, semantic entities are obtained. Comparing the results with the literatures shows that the results of this new approach can be sufficiently reliable.**

***Index Terms*—Disambiguation, Information Extraction, Semantic Entity Extraction, YAGO Ontology.**

I. Introduction

Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources such as texts. There are many systems to extract entities from a text. Each system extract their required entities from a text including Stanford named entities [1] and [2] named entities related biomedical [3] and terms in financial domain [4]. But none of them don't extract semantic entities, so they cannot be used for the applications that need to know semantic of entities such as computing semantic relatedness, semantic search and other works that need to semantic context.

For solving this problem, here a new system will be introduced called extracting semantic entities from texts. Semantic entities are the entities that their concepts are available in a knowledgebase. So, by extracting semantic entities from texts, an unstructured text space is converted into a structured semantic space. This extracting is done by a new disambiguation method that using YAGO ontology [5] as its knowledgebase that is a semantic space.

Disambiguation is a method that in which main sense of an ambiguous word in a text can be obtained. Disambiguation can be used for various aims. In this paper, this method is used to extract semantic entities from a text by introducing a new disambiguation method.

A knowledgebase can be the ontology, so the entities that are extracted by ontologies are semantical. Medelyan et al [9] claim the most appropriate work in this field is YAGO ontology. But ontologies only extract entities from structured texts such as infoboxs. In this paper, a new system is introduced to extract semantic entities from unstructured texts using YAGO as its knowledgebase.

Each one of previous disambiguation works disambiguate its ambiguous words, using a resource in which ambiguous words meaning and related knowledge are available. This resource is called "background knowledge". Bunescu et al [6], used encyclopedic knowledge as background knowledge. Mihalcea [7] and Sinha et al [8] used Wikipedia as background knowledge. But Medelyan et al [9] claim the most appropriate work in this field is YAGO ontology. For this reason, YAGO is used as the background knowledge of new disambiguation method. Since YAGO ontology has many semantic entities, so it can help to extract semantic entities from texts as a knowledgebase.

In previous works, Wikipedia was the best of background knowledge resource for disambiguation. Using Wikipedia as the background knowledge resource, in addition to its advantages, has two major problems. Firstly, Wikipedia is not completely reliable and then, information of this resource is textual and unstructured. Semantic information can't easily be extracted from unstructured resources. Suggestion of the present work can solve these problems. For this purpose, it is suggested that, instead of Wikipedia, YAGO ontology be used as background knowledge resource. Since YAGO ontology is obtained from Wikipedia, all its advantages are included. Besides, as YAGO ontology uses WordNet to prove its facts accuracy, so can be relied on. On the other hand, YAGO ontology is a structured knowledgebase, and a set of facts, which can be helpful in easily extracting semantic of entities. Each fact in ontology is as a triple that includes two entities and a relation between them. These triples can be used to extract entities from a text, obtain semantic of those entities.

The contributions of this paper are as follows:

- *Introducing a new method called semantic entity extraction.* Here, a new method is introduced to extract semantic entities from an unstructured text.
- *Introducing a new disambiguation method.* To extract semantic entities a new disambiguation method will be introduced that uses new background knowledge, and it will be shown that this background knowledge is state of the art for this paper purpose.
- *Creating a new application for YAGO ontology*. In this paper using YAGO as background knowledge is proposed and it will be shown that this ontology is

Manuscript received March 27, 2011. This work was supported in part by the Islamic Azad University, Roudsar and Amlash Branch.

Farhad Abedini is with the Electrical and Computer Engineering Department, and member of Young Researchers Club, Islamic Azad University, Roudsar and Amlash branch, Roudsar, Iran. Phone: +98-01426215051, e-mail: abedini.ac@gmail.com.

Fariborz Mahmoudi is with the Electrical and Computer Engineering Department, Islamic Azad University, Qazvin Branch, Qazvin, Iran. Phone: +9802813665275-3665276-, e-mail: mahmoudi@qiau.ac.ir.

Amir Hossein Jadidinejad is with the Electrical and Computer Engineering Department, Islamic Azad University, Qazvin Branch, Qazvin, Iran. Phone: +9802813665275-3665276, e-mail: amir@jadidi.info.

one of the most appropriate background knowledge resources for these aims.

- *Converting an unstructured text into a set of semantic entities*. The method that is introduced for semantic entity extraction can be used for converting an unstructured text into a set of semantic entities.

Thi*s* paper has been structured as follows. In next section first the solution for semantic entities extraction by new disambiguation method is described and then by using it, experimental results will be presented. These experimental results are performed on a benchmark dataset, introduced by Lee [10], and is compared with Stanford named entity recognition (NER), one of the best entity extraction systems. Finally, conclusions are represented.

II. Semantic Entity Extraction

Semantic entities are the entities that their concepts are available in a knowledgebase. Semantic entity extraction is a new method that is introduced in this paper. The solution for semantic entity extraction from a text by the new disambiguation method will be described as follows. First, the text must be preprocessing to be obtained unique string called tokens. Next, each of these tokens must be disambiguated using YAGO until in final semantic entities be obtained from text.

A. Preprocessing

Before semantic entity extraction by disambiguation method, the text must be preprocessed. Since characters, dates and numbers of the text can be an entity, so they can be considered as a semantic entity to be extracted from a text. But each of them can be in different forms to express its purposes. For example, “May 5th, 1983” and “1983-5-5” have a same meaning. So they should have a same structure to present a unique meaning. This work is done by normalization of them.

Different sources come with different encodings. But to have a unique meaning for the same contexts, a unique encoding must be used and other encodings must be changed into it. Here a method is introduced that converts all types of encodings into Unicode. For dates, ISO 6008 format is used and for numbers all of units are converted into SI units. End step of text normalization is to eliminate additional part of sentences. A same work in this field has been done in LEILA [11], and its idea has been used in this paper.

Then the text must be divided into small strings known as “tokens”. Here the method of SOFIE [12] is used to do this. In this method, a text is given as input and output is a set of tokens with their types.

Assigning each string into one of the token types, types of strings are specified. So unnecessary strings can be ignored and deleted. Now it must be shown that which of tokens can be semantic entities. For this reason, the next part proceeds on finding entities from obtained tokens.

B. New Disambiguation Method

YAGO ontology is a knowledgebase with high coverage and precision that has been obtained from Wikipedia and WordNet [5]. In fact, it can be said that it is state of the art of knowledge resources in mining meaning domain [9]. It contains about 2 million entities and 19 million facts about them and has only 99 unique relations. Previous ontologies had not this property. In YAGO, since only 99 unique relations are exist, so it is possible that same sentences that are explained with different forms can be mapped with one unique relation. For example, both *Born* and *Birthday* map to the relation *birthDate*. This is a good advantage for YAGO to be benefit for extracting semantic entities, because same concepts have only one unique form. So, the YAGO can be appropriate background knowledge for goal of semantic entity extraction. The entities of YAGO are completely semantical, because all relations of YAGO’s entities with each other are available. So each of tokens can be matched with one of YAGO entities, one can deduce that a semantic entity has been extracted. Here, this matching is introduced as “token disambiguation”.

There are many methods to disambiguate an ambiguous word. In previous works such as [1] disambiguation was used for entity extraction. But here disambiguation is used to extract semantic entity. For this aim in this paper, token is considered as an ambiguate word that can be classified in three statuses.

First, if it cannot be matched with YAGO entities, in consequence it is not desired entity and will be ignored. Second, if it can be matched only with one of YAGO entities, in consequence desired entity is found easily. And third, if it can be matched with several YAGO entities, in consequence the token is disambiguated with the method that comes in continue.

This method must select one of the matched entities as the semantic entity. For this aim matched entities are considered as different meanings of the token (ambiguate word). These different meanings are shown with e_i.

Then all of tokens that obtained from text are matched with YAGO entities. A set of YAGO entities is obtained. This set is shown with *e_set(t)* that t is text name.

Each of YAGO entities that is related with e_i in YAGO ontology, store in *e_set*(e_i).

Then intersection between all values of *e_set*(e_i) and *e_set(t)* must be compute. Number of relationships of each e_i with the text entities is shown with $|e\_set(t) \cap e\_set(e_i)|$.

Each of e_i (meanings of ambiguate token) that have more relationship with the text entities is more near to the text and can be resulted that this entity is main meaning of ambiguate token. In fact, the ambiguate token that was matched with several entities have been disambiguated. And nearest entity is obtained depending on the text. This token disambiguation method is shown in algorithm (1).

The inputs of this algorithm are a *token* that is obtained from preprocessing step that has been matched with several YAGO entities (the matching step comes in next algorithm called semantic entity extraction), a *text* that the token is extracted from there, the *YAGO ontology* for obtaining *e_set(t)* and *e_set*(e_i), and last input is *a set of entities* in YAGO ontology that are matched with the token and have been shown with e_i. The e_i comes from next algorithm called semantic entity extraction algorithm.

The output of algorithm (1) is only one semantic entity that is used for the semantic entity extraction algorithm. This means that one of the different meanings of token must be selected as semantic entity.

ALGORITHM TOKEN DISAMBIGUATION

Input: Token *token*, Text *t*, YAGO_Ontology *o*, Entities e_i

Output: Semantic Entity for *token*

```
e_set(t) := set of matched entities in o with all tokens in t
n: Number of e_i
FOR i = 1 TO n
    e_set(e_i) := set of entities related to e_i in o
FOR i = 1 TO n
    Number[i] := |e_set(t) ∩ e_set(e_i)|
FOR i = 1 TO n
    IF (Number[i] = Max) THEN  RETURN e_i
```

(1)

C. New Semantic Entity Extraction Algorithm

In previous part, it is shown how an ambiguous token can be disambiguated. In this part, this disambiguation algorithm is used to extract semantic entities from a text. All of steps that were introduced in this paper have been coming in algorithm(2).

The inputs of this algorithm are a *text* that semantic entities must be extracted from there, and the *YAGO ontology* as resource of semantic entities.

The output of this algorithm is a set of semantic entities that are extracted from the text.

ALGORITHM SEMANTIC ENTITY EXTRACTION

Input: Text *t*, YAGO_Ontology *o*

Output: A Set of Semantic Entities *se_set*

```
1 Preprocessing(t)
2 tokens(i) := set of tokens
3 m := numbers of tokens
4 FOR i =1 TO m
5    {
6       IF (Match tokens(i) with the entities in o) THEN
7          e_1,..,e_n := all of matched entities in o with tokens(i)
8       ELSE Continue
9       IF (n=1) THEN   se_set(i) := e_1
10      ELSE
11         se_set(i) := DISAMBIGUATION(tokens(i),t,o, e_i)
12   }
13 RETURN  se_set(i)
```

(2)

The preprocessing step is done in line 1. In this step, first the text is normalized and then the normalized text is divided into a set of tokens. In line 2, each of the tokens are assigned to *tokens(i)*. Numbers of these tokens that are extracted from the text are shown with *m* variable in the algorithm (line3).

The matching step is done in line 6 and 7. All entities in YAGO ontology that have been matched with *tokens(i)* are assigned in e_1 to ,e_n. In line 9, numbers of matched entities is checked. If numbers of matched entities be only one, then it can be resulted that semantic entity has been obtained easily. But if numbers of matched entities be more than one, then these matched entities must be disambiguated with algorithm (1). When algorithm (1) is called in line 11 with different meanings of the token (matched entities), then after executing algorithm (1), desired semantic entity is obtained. These operations are repeated for all text tokens until all semantic entities be obtained. Finally, in the last line a set of semantic entities must be returned.

So by this method each of tokens can be matched with one of YAGO entities. Since this ontology is a knowledgebase and its information can be relied (with more than 95% confidence) also each of entities in YAGO has certain relations [5], so it can be claimed that the *semantic entities* have been obtained.

All of steps to extract semantic entities from a text are shown in figure1. In this figure converting an unstructured text into a set of structured semantic entities is cleared.

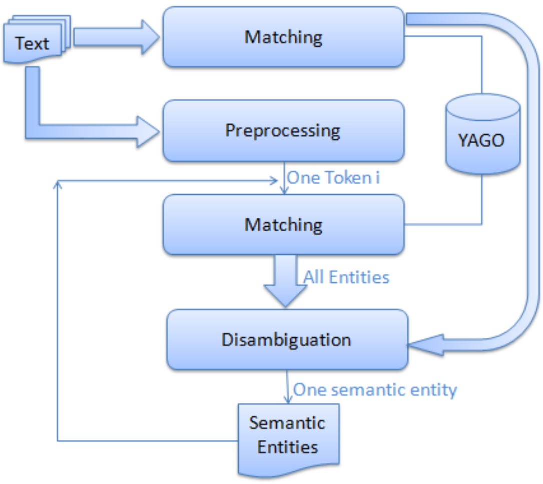


Figure 1. Converting an unstructured text into a set of semantic entities

In figure1, it is shown how an unstructured text space is converted into a structured semantic space by extracting semantic entities from texts. YAGO ontology is used for matching words or tokens of text with entities that are exist in YAGO ontology. In disambiguation step, one of matched entities is selected as semantic entity for a token.

III. Experimental Result

A. Implementation

To implement this project, first YAGO ontology was converted into Mysql database. This work was performed by a computer with 2G RAM and CPU Dual Core with 3M Cache. Its runtime took 22 days. The result was a database of triple facts. Its properties have been shown in table 1.

Steps of preprocessing, and two algorithms of disambiguation and semantic entity extraction, have been implemented with java codes on this database.

TABLE I. YAGO Properties In Mysql Implementation

Table Name	Data length	Index length	Fields	#row(million)
Entities	114.6 MB	0	Name	2
Facts	2.6 GB	12 GB	Relation, Arg1, Arg2	19

In table 2 it is shown that YAGO ontology has been converted into a database with two tables. An entity table that is contains about 2 million semantic entities, and a facts table that is contains about 19 million triple about entities called facts. In each of the facts there are two entities and a relation between them. These facts exist in the real world. This property of facts is very helpful for extracting semantic entities from the text.

In table 2 it is shown that YAGO ontology has been converted into a database with two tables. An entity table that is contains about 2 million semantic entities, and a facts table that is contains about 19 million triple about entities called facts. In each of the facts there are two entities and a relation between them. These facts exist in the real world. This property of facts is very helpful for extracting semantic entities from the text.

B. Evaluation

To evaluate semantic entity extraction method that was presented in this paper, this method is compared with NER one of the best named entity recognition that is implemented by Stanford Natural Language Processing Group [1].

In this work the Lee benchmark dataset [10], is used, because the authors are going to work on this datasets in future works for computing semantic relatedness of texts. Also, state of the art of computing semantic relatedness has been introduced in [16] and [17]. This dataset contains a collection of 50 documents from the Australian Broadcasting Corporation's news mail service. This datasets have given to some peoples and have requested them to find all semantic entities in these documents. To compare our work with NER, this judgment is used. This means that each of NER or our work is measured with this judgment. And the result of that is shown in table 1.

TABLE II. RESULT OF NER AND SESR COMPARISON

	Recall	Precision
Semantic Entity Extraction	95%	98%
NER	90%	90%

The result of table 2 can be shown in figure 2 otherwise.

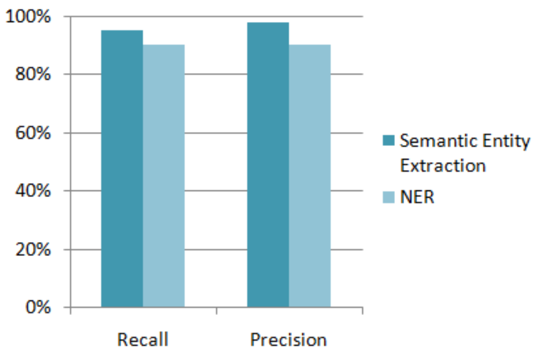


Figure 2. Recall and Precision of NER and Semantic Entity Extraction.

Precision and recall of NER and semantic entity extraction method was compared with the human judgments. The results show that semantic entity extraction method can lead to more accurate results on this dataset. For a case study the three texts from the dataset was selected that have been shown in table 2. The results of entity extraction have been shown in table 3.

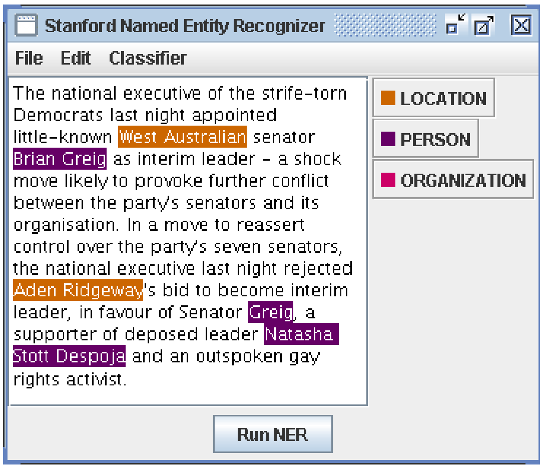
TABLE III. THREE TEXTS FOR CASE STUDY

#Txt	Text
1	The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven senators, the national executive last night rejected Aden Ridgeway's bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist.
2	Cash-strapped financial services group AMP has shelved a $400 million plan to buy shares back from investors and will raise $750 million in fresh capital after profits crashed in the six months to June 30. Chief executive Paul Batchelor said the result was "solid" in what he described as the worst conditions for stock markets in 20 years. AMP's half-year profit sank 25 per cent to $303 million, or 27c a share, as Australia's largest investor and fund manager failed to hit projected 5 per cent earnings growth targets and was battered by falling returns on share markets.
3	The United States government has said it wants to see President Robert Mugabe removed from power and that it is working with the Zimbabwean opposition to bring about a change of administration. As scores of white farmers went into hiding to escape a round-up by Zimbabwean police, a senior Bush administration official called Mr Mugabe's rule "illegitimate and irrational" and said that his re-election as president in March was won through fraud. Walter Kansteiner, the assistant secretary of state for African affairs, went on to blame Mr Mugabe's policies for contributing to the threat of famine in Zimbabwe.

TABLE IV. COMPARING EXTRACTED ENTITIES BY TWO METHOD

#Txt	NER	Semantic Entity Extraction
1	LOCATION: West Australian Aden Ridgeway PERSON: Brian Greig Greig Natasha Stott Despoja	West_Australian Brian_Greig Number: 7 Aden_Ridgeway Natasha_Stott_Despoja
2	LOCATION: Australia PERSON: Paul Batchelor ORGANIZATION: AMP	Numbers: 400000000#dollar, 750000000#dollar, 6, -06, -30, 20, 25, 27, 5, 303000000#dollar Australia Paul _Batchelor AMP
3	LOCATION: United States Zimbabwean African Zimbabwe PERSON: Robert Mugabe Mr Mugabe Walter Kansteiner	United_States Robert_Mugabe Zimbabwe Walter_H._Kansteiner,_III Africa

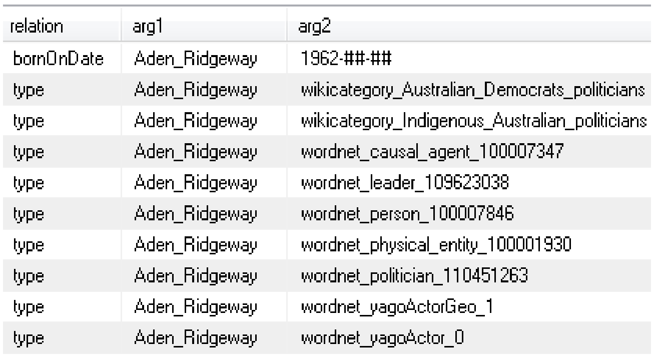
The results in table 2 show difference between NER method and semantic entity extraction method. Almost, all of entities in NER are extracted in semantic entity extraction method. And semantic entity extraction method has more entities than NER. One of advantage of semantic entity extraction method is that repetitive entities are not available, and there is only one form of them. But in NER for example in text1 there is two Greig. This problem has been come in figure3.



The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven senators, the national executive last night rejected Aden Ridgeway's bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist.

Figure 3. Entities in text1 that have been extracted by NER

One of disadvantage of NER is that some entities are extracted wrong. For example in figure 3 "Aden Ridgeway" was extracted as location. But it is clear that it is a person. But as is clear in figure 4, in semantic entity extraction method this entity has been extracted correctly.



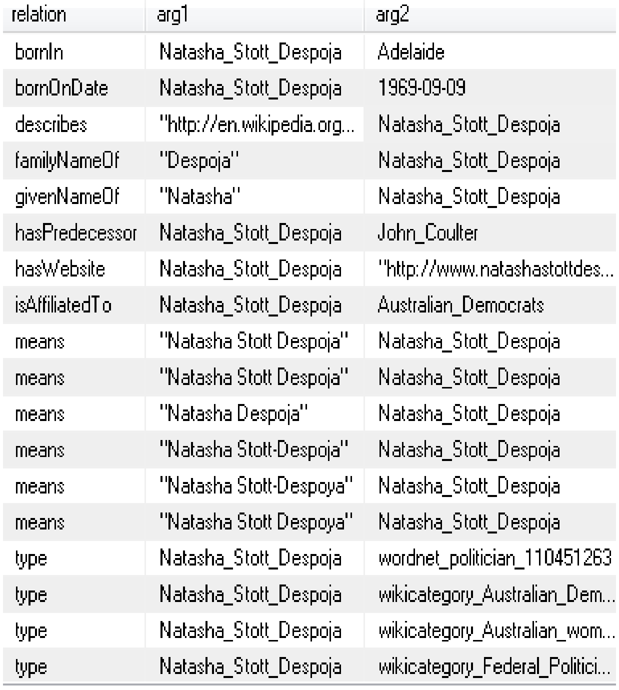
relation	arg1	arg2
bornOnDate	Aden_Ridgeway	1962-##-##
type	Aden_Ridgeway	wikicategory_Australian_Democrats_politicians
type	Aden_Ridgeway	wikicategory_Indigenous_Australian_politicians
type	Aden_Ridgeway	wordnet_causal_agent_100007347
type	Aden_Ridgeway	wordnet_leader_109623038
type	Aden_Ridgeway	wordnet_person_100007846
type	Aden_Ridgeway	wordnet_physical_entity_100001930
type	Aden_Ridgeway	wordnet_politician_110451263
type	Aden_Ridgeway	wordnet_yagoActorGeo_1
type	Aden_Ridgeway	wordnet_yagoActor_0

Figure 4. Relations of Aden_Ridgeway entity

Relations of Aden_Ridgeway entity that extracted by semantic entity extraction method from text1, have been shown in figure 4. The relation of *bornOnDate* for this entity shows that this entity is a person that has been born on 1962. Other relations define other concepts and facts about this person in real world. So semantic of entities in method of semantic entity extraction completely are available.

So it can be seen in table 3, the semantic entity extraction method is more benefit than NER. NER does not extract semantic entities and gives only type of entities whereas in the semantic entity extraction method entities have matched with synonymous entities in YAGO. In this method, type of entity obtained in token extraction step. Since the YAGO entities are completely semantical, so it can be claimed that the entities which obtained with this method are "semantic entities". For example, some of facts about one of entities (Natasha_Stott_Despoja) that extracted by our method are shown in figure 5. So it can be resulted that this entity is semantical.

In figure5, it is shown that the extracted entity that has been obtained by this paper method exist in YAGO ontology. In fact, it is one of YAGO entities. So, all of its existence relations in YAGO with another entities are available. Each row of these triples (relation, entity1, entity2) forms a fact. Some of these facts have been shown in figure5. For example, in triple of "bornIn, Natasha_Stott_Despoja, Adelaide" there is a fact in real world that say the Natasha Stott Despoja has been born in Adelaide. In this fact there are two entities (Natasha_Stott_Despoja and Adelaide) and one relation between them.



relation	arg1	arg2
bornIn	Natasha_Stott_Despoja	Adelaide
bornOnDate	Natasha_Stott_Despoja	1969-09-09
describes	"http://en.wikipedia.org...	Natasha_Stott_Despoja
familyNameOf	"Despoja"	Natasha_Stott_Despoja
givenNameOf	"Natasha"	Natasha_Stott_Despoja
hasPredecessor	Natasha_Stott_Despoja	John_Coulter
hasWebsite	Natasha_Stott_Despoja	"http://www.natashastottdes...
isAffiliatedTo	Natasha_Stott_Despoja	Australian_Democrats
means	"Natasha Stott Despoja"	Natasha_Stott_Despoja
means	"Natasha Stott Despoja"	Natasha_Stott_Despoja
means	"Natasha Despoja"	Natasha_Stott_Despoja
means	"Natasha Stott-Despoja"	Natasha_Stott_Despoja
means	"Natasha Stott-Despoya"	Natasha_Stott_Despoja
means	"Natasha Stott Despoya"	Natasha_Stott_Despoja
type	Natasha_Stott_Despoja	wordnet_politician_110451263
type	Natasha_Stott_Despoja	wikicategory_Australian_Dem...
type	Natasha_Stott_Despoja	wikicategory_Australian_wom...
type	Natasha_Stott_Despoja	wikicategory_Federal_Politici...

Figure 5. Some facts of Natasha_Stott_Despoja entity in YAGO.

For each entity there are many relations in YAGO that explain the facts about it. So these entities are certainly semantic entities. In extracting semantic entities from Lee dataset, numbers of frequency of each relation have been obtained and it has been shown in table 5 and figure 6.

TABLE V. FREQUENCY OF RELATIONS BY OFE ON LEE DATASET

Relation	Domain	Range	%Facts
means	yagoWord	entity	26.73
type	entity	yagoClass	22.53
inLanguage	yagoFact	language	17.81
isCalled	entity	yagoWord	10.93
describes	yagoURL	entity	10.62
familyNameOf	yagoWord	person	2.85
givenNameOf	yagoWord	person	2.84
bornOnDate	person	yagoDate	2.21
subClassOf	yagoClass	yagoClass	1.25
diedOnDate	person	yagoDate	1.03
Other			1.2

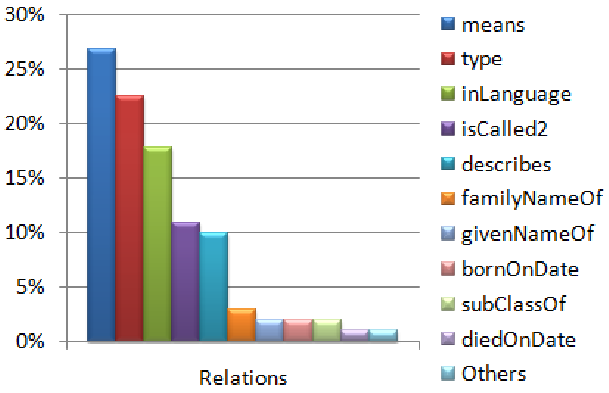
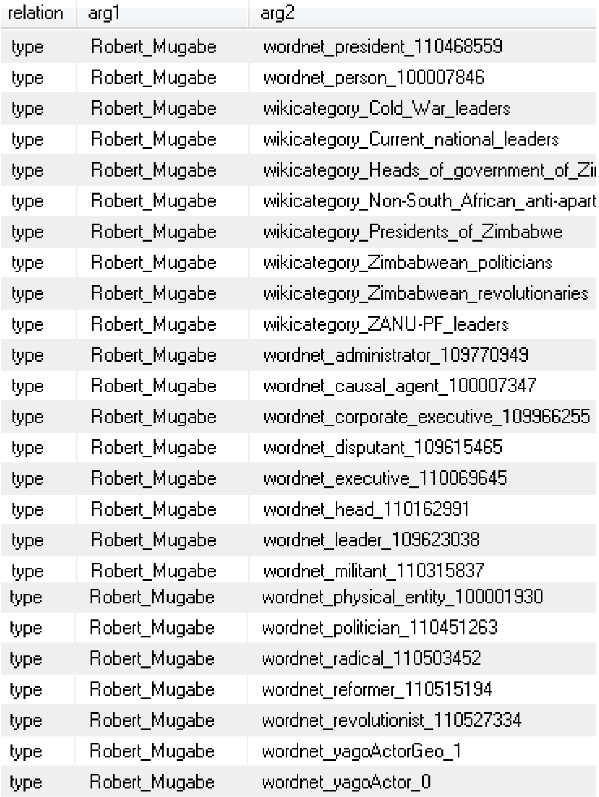


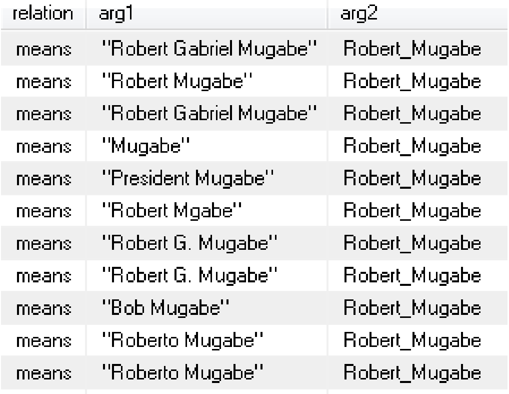
Figure 6. Frequency of YAGO relations.

In figure 6 it is clear that two relation means and type have most frequency. In [5] it has been come that these relations there are for any entity in YAGO. This subject can be very helpful in some problems such as computing semantic relatedness. This problem is performed in [13] and [14] by author. For example two relations type and means have been shown in figures 7 and 8 for the entity of Robert_Mugabe from the case study text 3.



relation	arg1	arg2
type	Robert_Mugabe	wordnet_president_110468559
type	Robert_Mugabe	wordnet_person_100007846
type	Robert_Mugabe	wikicategory_Cold_War_leaders
type	Robert_Mugabe	wikicategory_Current_national_leaders
type	Robert_Mugabe	wikicategory_Heads_of_government_of_Zi
type	Robert_Mugabe	wikicategory_Non-South_African_anti-apart
type	Robert_Mugabe	wikicategory_Presidents_of_Zimbabwe
type	Robert_Mugabe	wikicategory_Zimbabwean_politicians
type	Robert_Mugabe	wikicategory_Zimbabwean_revolutionaries
type	Robert_Mugabe	wikicategory_ZANU-PF_leaders
type	Robert_Mugabe	wordnet_administrator_109770949
type	Robert_Mugabe	wordnet_causal_agent_100007347
type	Robert_Mugabe	wordnet_corporate_executive_109966255
type	Robert_Mugabe	wordnet_disputant_109615465
type	Robert_Mugabe	wordnet_executive_110069645
type	Robert_Mugabe	wordnet_head_110162991
type	Robert_Mugabe	wordnet_leader_109623038
type	Robert_Mugabe	wordnet_militant_110315837
type	Robert_Mugabe	wordnet_physical_entity_100001930
type	Robert_Mugabe	wordnet_politician_110451263
type	Robert_Mugabe	wordnet_radical_110503452
type	Robert_Mugabe	wordnet_reformer_110515194
type	Robert_Mugabe	wordnet_revolutionist_110527334
type	Robert_Mugabe	wordnet_yagoActorGeo_1
type	Robert_Mugabe	wordnet_yagoActor_0

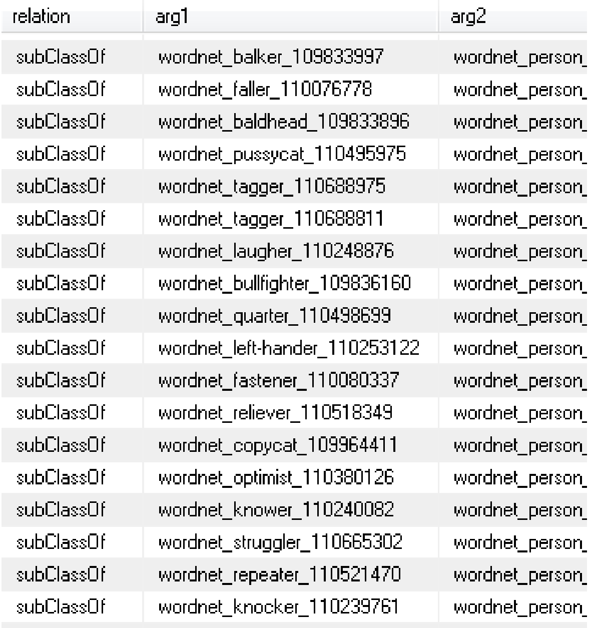
Figure 7. Facts of TYPE relaition in YAGO.



relation	arg1	arg2
means	"Robert Gabriel Mugabe"	Robert_Mugabe
means	"Robert Mugabe"	Robert_Mugabe
means	"Robert Gabriel Mugabe"	Robert_Mugabe
means	"Mugabe"	Robert_Mugabe
means	"President Mugabe"	Robert_Mugabe
means	"Robert Mgabe"	Robert_Mugabe
means	"Robert G. Mugabe"	Robert_Mugabe
means	"Robert G. Mugabe"	Robert_Mugabe
means	"Bob Mugabe"	Robert_Mugabe
means	"Roberto Mugabe"	Robert_Mugabe
means	"Roberto Mugabe"	Robert_Mugabe

Figure 8. Facts of MEANS relaition in YAGO.

Studies have shown that although the *means* relation is most relation in extracted facts, but the *type* relation is more benefit to most applications. In these figures, it is clear that the *type* relation is more structured and more useful than *means*. In table 5 it was shown that domain of the *type* relation is entity that extract by semantic entity extraction method, and range of the *type* relation is *yagoClass* that gives upper class of this entity. Having upper class of class by the *subClassOf* relation that available in table 5, upper context of entity will be obtained that help us to solve many problems that need to *Is_A* relationships easily. In figure 9 some of facts about *subClassOf* relation have been shown. In this figure subclass of person class is shown.



relation	arg1	arg2
subClassOf	wordnet_balker_109833997	wordnet_person_
subClassOf	wordnet_faller_110076778	wordnet_person_
subClassOf	wordnet_baldhead_109833896	wordnet_person_
subClassOf	wordnet_pussycat_110495975	wordnet_person_
subClassOf	wordnet_tagger_110688975	wordnet_person_
subClassOf	wordnet_tagger_110688811	wordnet_person_
subClassOf	wordnet_laugher_110248876	wordnet_person_
subClassOf	wordnet_bullfighter_109836160	wordnet_person_
subClassOf	wordnet_quarter_110498699	wordnet_person_
subClassOf	wordnet_left-hander_110253122	wordnet_person_
subClassOf	wordnet_fastener_110080337	wordnet_person_
subClassOf	wordnet_reliever_110518349	wordnet_person_
subClassOf	wordnet_copycat_109964411	wordnet_person_
subClassOf	wordnet_optimist_110380126	wordnet_person_
subClassOf	wordnet_knower_110240082	wordnet_person_
subClassOf	wordnet_struggler_110665302	wordnet_person_
subClassOf	wordnet_repeater_110521470	wordnet_person_
subClassOf	wordnet_knocker_110239761	wordnet_person_

Figure 9. Some Facts of subClassOf relaition in YAGO.

C. Limitations

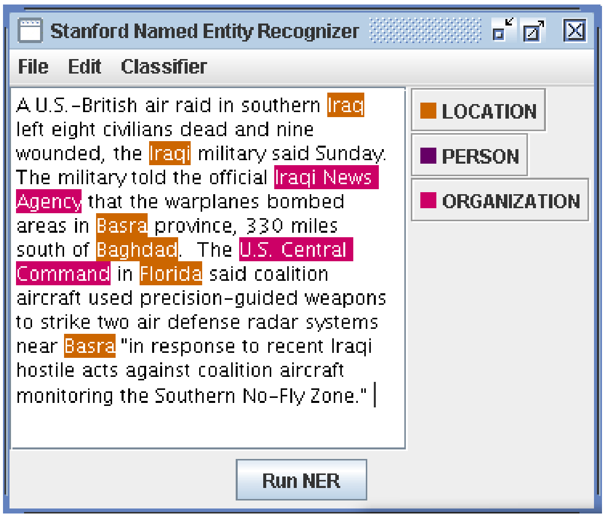
The semantic entity extraction method that was introduced in this paper, besides its good advantage, has the some limitations that will be explained. Since, entities in YAGO ontology are limited, and since semantic entities are extracted from YAGO, so the semantic entities that are extracted from the text by this method are limited. Although the YAGO entities are very much (about 2 million), but the entities may be that this method cannot extract them. But the experimental results on Lee dataset show that these entities are very little (about 5%). This limitation in NER was 10%. This problem is shown in table 6.

TABLE VI. LIMITATION OF ENTITY EXTRACION METHODS

Method	Limitation
Semantic Entity Extraction	5%
NER	10%

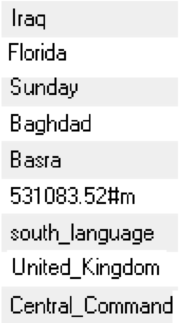
For example, in figure 10 it is clear that some entities in semantic entity extraction method are available that there are not in NER such as Sunday and United Kingdom. It shows advantage of semantic entity extraction method. Or some entities are extracted wrong in NER such as Aden Ridgeway in figure 3.

As previously mentioned, there are few limitations in semantic entity extraction method. For example in figure 11, semantic entities of text11 have been shown. The entity of “Iraqi News Agency” is not available, because this entity is not in YAGO ontology. So, if the entity is not in YAGO ontology, then it can be said that the method of semantic entity extraction cannot extract it, and this is limitation of this method.



A U.S.-British air raid in southern Iraq left eight civilians dead and nine wounded, the Iraqi military said Sunday. The military told the official Iraqi News Agency that the warplanes bombed areas in Basra province, 330 miles south of Baghdad. The U.S. Central Command in Florida said coalition aircraft used precision-guided weapons to strike two air defense radar systems near Basra "in response to recent Iraqi hostile acts against coalition aircraft monitoring the Southern No-Fly Zone."

Figure 10. Entities in text10 that have been extracted by NER



Iraq
Florida
Sunday
Baghdad
Basra
531083.52#m
south_language
United_Kingdom
Central_Command

Figure 11. Entities in text10.

Recently, YAGO2 [15] is created. This ontology has 10 million entities instead of previous 2 million. This means that the limitation will be improved by YAGO2. For this reason in future work of authors, this improving is done.

IV. Conclusion And Future Works

In this paper, the approach of extracting semantic entities from a text by new disambiguation method that using YAGO ontology was presented. In evaluation it was shown that this method is benefit to extract semantic entities.

The contributions of this paper was introducing a new semantic entity extraction method, introducing a new disambiguation method, creating a new application for YAGO ontology, and converting an unstructured text into a set of semantic entities.

As mentioned in experimental results, all of entities that are extracted by our method, their facts in real world are available in YAGO. It means that these entities are completely semantical.

In our next work we are going to use method of semantic entity extraction to compute semantic relatedness of texts. We consider using some YAGO relations such as *means* and *type* to find upper context for computing semantic relatedness. These relations are available for all entities in YAGO ontology.

Since relations between YAGO entities are available in YAGO ontology, we also consider using semantic entities that was obtained from our method, to extract facts from text. These facts can be used for computing semantic relatedness between texts. As previously mentioned, recently a newer version of YAGO called YAGO2 has been created that is very more complete. To improving semantic entity extraction method and increase of its limitations, we are going to use this ontology in our future works.

References

[1] The Stanford Natural Language Processing Group. *Stanford Named Entity Recognizer (NER)*, version 1.1.1, 2009-01-16, http://nlp.stanford.edu/software

[2] Finkel, J.R., Manning, C.D.: *Joint parsing and named entity recognition*. In: North American Association of Computational Linguistics (NAACL), 2009.

[3] Alex, B., B. Haddow, and C. Grover, *Recognising Nested Named Entities in Biomedical Text*, in BioNLP 2007: Biological, translational, and clinical language processing. 2007: Prague.

[4] Feiyu Xu, Daniela Kurz, Jakub Piskorski, Sven Schmeier : *Term Extraction and Mining of Term Relations from Unrestricted Texts in the Financial Domain*, In: 5th International Conference on Business Information Systems; Poznan, Poland, 2002.

[5] F. M. Suchanek, G. Kasneci, and G. Weikum. *YAGO - A Large Ontology from Wikipedia and WordNet*. Elsevier Journal of Web Semantics, 6(3):203-217, September 2008.

[6] Bunescu, Razvan & Marius Pas¸ca. Using encyclopedic knowledge for named entity disambiguation. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 3–7 April 2006, pp. 9–16.

[7] Rada Mihalcea, Using Wikipedia for Automatic Word Sense Disambiguation, in Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007), Rochester, April 2007.

[8] Sinha, R. and Mihalcea, R., Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In Proc. of ICSC, 2007.

[9] Olena Medelyan, David Milne, Catherine Legg, Ian H.Witten. *Mining meaning from Wikipedia*. Elsevier Journal of Human-Computer Studies, Pages 716–754, May 2009.

[10] Michael D. Lee, Brandon Pincombe, and Matthew Welsh. *An empirical evaluation of models of text documents similarity*. In CogSci2005, pages 1254–1259, 2005.

[11] F. M. Suchanek, G. Ifrim, and G. Weikum. *LEILA: Learning to extract information by linguistic analysis*. In P. Buitelaar, P. Cimiano, and B. Loos, editors, Proceedings of the 2nd Workshop on Ontology Learning and Population (OLP2) at COLING/ACL 2006, pages 18–25, Sydney, Australia, 2006. Association for Computational Linguistics.

[12] F. M. Suchanek, M. Sozio, G. Weikum. *SOFIE: a self-organizing framework for information extraction*. In: Proceedings of the WWW 2009 conference ,2009.

[13] Farhad Abedini, Fariborz mahmoudi, Amir Hossein Jadidinejad. *A New Disambiguation Method for Semantic Entity Extraction Using YAGO Ontology*. In proceedings of IEEE 3th International Conference on Machine Learning and Computing (ICMLC 2011), Singapore, 26-28 February 2011, pp 79-83.

[14] Farhad Abedini, Fariborz mahmoudi, Amir Hossein Jadidinejad. *OFE: Ontological Facts Extraction from text for computing Semantic Relatedness*. In proceedings of IEEE 3th International Conference on Machine Learning and Computing (ICMLC 2011), Singapore, 26-28 February 2011, pp 84-88.

[15] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Gerhard Weikum.YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Research Report MPI-I-2010-5-007, Max-Planck-Institut für Informatik, November 2010.

[16] Ofer Egozi, Shaul Markoitch, and Evgeniy Gabrilovich. "Concept-Based Information Retrieval using Explicit Semantic Analysis", ACM Transactions on Information Systems, Vol. 0, No. 0, 2000, Pages 1–38, 2010.

[17] Evgeniy Gabrilovich and Shaul Markovitch, "Wikipedia-based Semantic Interpretation for Natural Language Processing", Journal of Artificial Intelligence Research 34 (2009) 443-498.