Entity Disambiguation in Text by YAGO Ontology

Farhad Abedini and Seyedeh Masoumeh Mirhashem

***Abstract*—Word disambiguation has been used for various applications. In this paper, disambiguation is used for the new semantic entity extraction system that will be suggested here. For this aim, a new disambiguation method will be introduced. To extract a semantic entity, a background knowledge resource must be used. For this disambiguation, YAGO ontology is used as the state-of-the-art of background knowledge in this field. Since entities in YAGO are meaningful, so in this disambiguation, semantic entities are obtained.**

***Index Terms*—Disambiguation, semantic entity extraction, YAGO ontology, background knowledge.**

I. Introduction

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. Semantic entity extraction can be used for many applications such as computing semantic relatedness, semantic search and other works that need to semantic context.

There are many systems to extract entities from a text. Each one of systems extract their required entities from a text including Stanford named entities [1]-[2] named entities related biomedical [3] and terms in financial domain [4]. But here, a system will be introduced to extract semantic entities, by a new disambiguation method that using YAGO ontology [5] as its background knowledge resource.

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. The entities that are extracted by ontologies are semantical. But ontologies only extract entities from structured texts such as infoboxs. In this disambiguation method, YAGO ontology is used to extract semantic entities from unstructured texts.

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 disambiguation method.* In this method, new background knowledge is used. And it will be shown that this background knowledge is most appropriate for this paper purpose.
- *Introducing a new entity extraction method.* Here, a new method is introduced to extract semantic entities from a text, using a *new disambiguation method.*
- *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 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.

This 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

The solution for semantic entities extraction from a text by new disambiguation method can be described as follows.

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

Manuscript received September 10, 2012; revised December 10, 2012. 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 a member of Young Researchers Club, Islamic Azad University, Roudsar and Amlash branch, Phone: +98-01426215051, Roudsar, Iran. (e-mail: abedini.ac@gmail.com)

Seyedeh Masoumeh Mirhashem is with the Islamic Azad University, Roudsar and Amlash branch, Phone: +98-01426215051, Roudsar, Iran. (e-mail: saeedeh.mirhashem@yahoo.com)

DOI: 10.7763/IJCTE.2013.V5.724

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 the most appropriate available knowledge resource in mining meaning domain [9]. It contains about 2 million entities and 15 million facts about them and has only 99 unique relations. So it can be appropriate background knowledge for this goal. The entities of YAGO, since all relations of YAGO’s entities with each other are available, are completely semantical. 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 is considered as different meaning of token (ambiguate word). These different meaning is shown with ei. 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 ei in YAGO ontology, store in e_set(ei). Then intersection between all values of e_set(ei) and e_set(t) must be compute. Number of relationships of each ei with the text entities is shown with |e_set(t)∩e_set(ei)|. Each of ei (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).

```
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 $e_i$
```

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).

```
Algorithm Semantic Entity Extraction
  Input:Text t, YAGO_Ontology o
  Output:A Set of Semantic Entities se_set
  tokens(i) : set of tokens
  tokens(i) := Preprocessing(t)
  m:= number of tokens
  FOR i =1 TO m
     {
  IF (Match tokens(i) with the entities in o) THEN
  $e_1$,..,$e_n$ := all of matched entities in o with tokens(i)
  ELSE Continue
  IF (n=1) THEN   se_set(i) := $e_1$
  ELSE
  se_set(i) := DISAMBIGUATION(tokens(i),t,o, $e_i$)
     }
  RETURN  se_set(i)
```

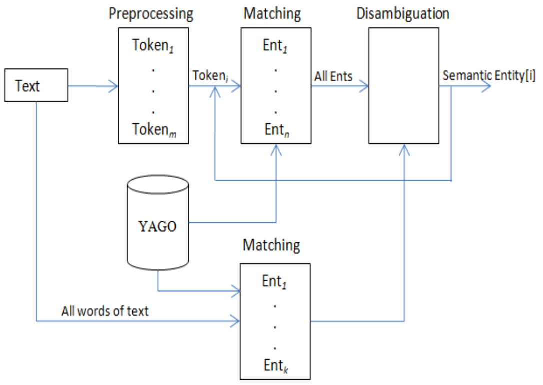


Fig. 1. Model of semantic entity extraction method.

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 entity in YAGO has certain relation [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 Fig. 1.

In Fig. 1, it is shown how 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 semanticcentity for a token.

III. Experimental Result

A. Implementation

To implement this project, first YAGO ontology has been 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 with volume 4G.

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

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. 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 I.

TABLE I: Result of NER and SESR Comparison

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

Precision and recall of NER and semantic entity extraction method was compared with the. The results show that on this dataset semantic entity extraction method can lead to more accurate results for our purpose. For a case study the three texts from the dataset was selected that have been shown in Table II. The results of entity extraction have been shown in Table III.

It can be seen in table III, for our purpose in these texts our method is better than NER. NER does not extract semantic entities and gives only type of entities whereas in our 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 we can claim that the entities which obtained with our method are "*semantic entities*". For example, some of facts about one of entities (Natasha_Stott_Despoja) that extracted by our method are shown in Fig. 2. So it can be resulted that this entity is semantical.

TABLE II: Three Texts for Case Study

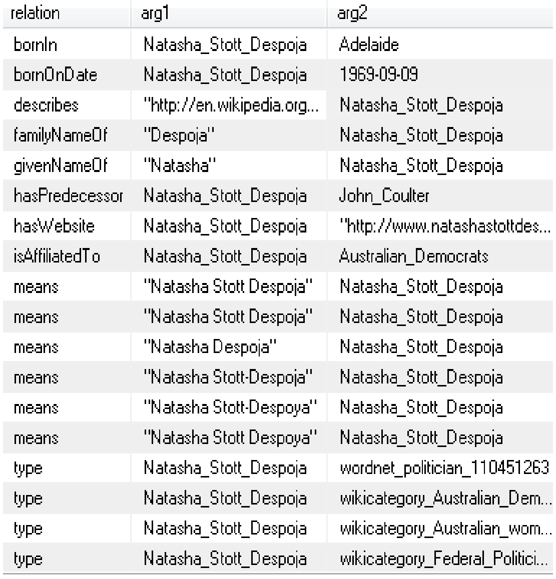
#Txt	Text
1	The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven 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 III: 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 - Bush - Mr Mugabe - Walter Kansteiner	United_States Robert_Mugabe Zimbabwe Walter_H._Kansteiner,_III Africa

In Fig. 2, 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.

These facts have been shown in Fig. 2.



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...

Fig. 2. Facts of an entity in YAGO

IV. Conclusion

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 our method is benefit to extract semantic extraction.

The contributions of this paper was introducing a new disambiguation method, introducing a new entity extraction 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 are available in YAGO. These facts explain all relations of entities with other entities. So it can be resulted that these entities are semantical.

Since, this method extracts semantic entities, so we can use to solve open problems such as semantic relatedness. The method introduced here can improve computing semantic relatedness. For this aim, in our next work we are going to use this method 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.

References

[1] The Stanford Natural Language Processing Group, *Stanford Named Entity Recognizer (NER)*, version 1.1.1, 16 Jan. 2009.

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

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

[4] F.-Y. Xu, D. Kurz, J. Piskorski, and S. Schmeier, "Term Extraction and Mining of Term Relations from Unrestricted Texts in the Financial Domain," presented at 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*, vol. 6, no. 3, pp. 203-217, September 2008.

[6] R. Bunescu and P. Ç. Marius, "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] R. Mihalcea, "Using Wikipedia for Automatic Word Sense Disambiguation," in *Proceedings of the North American Chapter of the Association for Computational Linguistics*, Rochester, April 2007.

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

[9] O. Medelyan, M. David, L. Catherine, and I. H.Witten. "Mining meaning from Wikipedia," *Elsevier Journal of Human-Computer Studies*, pp. 716–754, May 2009.

[10] D. L. Michael, B. Pincombe, and M. Welsh, "An empirical evaluation of models of text documents similarity," in *Proc. of CogSci2005*, pp. 1254–1259, 2005.

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

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

Farhad Abedini was born in Roudsar, Iran in 1983. He received B.E. degree in computer software engineering in 2008 from Tabarestan University Iran, M. Sc. degree in computer engineering in 2011 from Islamic Azad University, Qazvin Branch, Qazvin, Iran (QIAU). He has been a lecturer of computer engineering department at the Islamic Azad University, Roudsar and Amlash Branch, Roudsar, Iran (RAIAU). Now, he is head of Young Researchers Club of Roudsar Branch. He is also member of Young Researchers Club and IACSIT. He was secretary and reviewer of some conferences, seminars and academic competitions. His research interests include Information Retrieval, Text Mining, Question Answering Knowledge Extraction from huge scale Collaboratively Constructed Semantic Resources and using new methods of knowledge representation in semantic processing. CV of Farhad Abedini is available in http://abedini.org/EnPage/En.htm.

Seyedeh Masoumeh Mirhashem was born in Roudsar, Iran in 1983. She received B.A. degree in English language translator in 2011 from Payamnoor University (PNU) of Roudsar Branch, Roudsar, Iran. Now, she is student of M.A. degree in English language teaching at Payamnoor University (PNU) of Rasht Branch, Rasht, Iran. She is member of Young Researchers Club, Islamic Azad University, Roudsar and Amlash Branch, Roudsar, Iran. She also received post diploma in computer software in 2003. His research interests include Linguistics, Information Retrieval, Text Mining, Question Answering, and Motivation in Learning.