From Text to Facts: Recognizing Ontological Facts for a New Application

Farhad Abedini and Seyedeh Masoumeh Mirhashem

Abstract—**Fact extraction methods are used for various aims. Here, a new method is introduced for a new application – computing semantic relatedness of texts. Ontological facts are the facts about real world that are available in a knowledgebase called ontology. There are many systems to extract ontological facts from a text. State-of-the-art of these systems is SOFIE that is not appropriate to compute semantic relatedness of texts. In this paper, this problem will be explained and a new ontological facts extraction system is introduced that is appropriate for computing semantic relatedness of texts. For this aim, YAGO ontology will be used as a background knowledge and facts resource. In final it will be suggested that using YAGO relations can optimize ontological facts extraction system.**

Index Terms—**Ontological facts, semantic relatedness, disambiguation, semantic entity extraction, YAGO ontology.**

I. Introduction

Ontology is a knowledgebase in which a computer processable collection of knowledge is available about the world as collection of facts. Various ontologies have been tried to obtain desired ontological facts from available resources. State-of-the-art of these ontologies is YAGO [1]. YAGO uses Wikipedia as resource that is best available resource on the web about the world. In YAGO, a fact is shown by a triple of two entities and one relation between them.

YAGO obtains its ontological facts from structured text in Wikipedia such as Infoboxs. To extract ontological facts from unstructured text, SOFIE [2] was introduced. SOFIE is state-of-the-art for extracting ontological facts from unstructured text. Since, SOFIE extracts only the facts that their accuracy has been inference by SOFIE system, so it cannot extracts many facts from text.

In computing semantic relatedness, some texts must be compared together to find semantic relatedness between them. In this paper, it is suggested to convert each one of texts into a set of ontological facts and then these sets of facts are compared. As previously mentioned, SOFIE cannot extract many facts from the text, so it cannot benefit method to extract ontological facts for computing semantic relatedness. Therefore, it cannot helps to compute semantic relatedness of texts.

To solve this problem, in this paper a suggestion will be discussed. The suggestion comes in continue. First each one of the texts must be converted into a set of YAGO entities called “semantic entities”, and then by these entities and matching them with YAGO facts, desired ontological facts are obtained.

Previous works in the semantic relatedness area convert a text to a set of words and then measure the relatedness between these words and the words obtained from another text. These are not conceptual comparison. But in the present work, for the conceptual comparison it will be suggested that, all of the texts must be converted into a set of ontological facts and then sets of ontological facts of different texts be compared with each other. This comparison can be conceptual, because the fact of the world is compared together.

To extract ontological facts from a text, semantic entities must be extracted. To extract entities from a text, a resource is needed by which context of information can be extracted from text. This resource is called “background knowledge”. For example, in [3] lexical databases such as WordNet [4] and Roget’s Thesaurus [5] were used as background knowledge. Creation of lexical resources requires lexicographic expertise, and takes a lot of time and effort. To solve this problem, Corpus-based approaches obtain its background knowledge by performing statistical analysis of large untagged document collections. The most successful and well known of these techniques is Latent Semantic Analysis (LSA) [6], which relies on the tendency for related words to appear in similar contexts. But it can only provide accurate judgments when the corpus is very large, and consequently the pre processing effort required is significant. To solve this problem, Explicit Semantic Analysis (ESA) was introduced. Gabrilovich et al [7] showed that ESA is stat-of-the-art in computing texts semantic relatedness, in which Wikipedia has been used as background knowledge. Such background knowledge must have some properties. One of the most important properties is that in the background knowledge, information about every possible thing should be existed in the world. It is clear that such a resource is not available. But as proved by Medelyan et al [8], Wikipedia is the most appropriate existent resource in this field.

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

Manuscript received April, 2012; revised May 30, 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 member of Young Researchers Club, Islamic Azad University, Roudsar and Amlash branch, Roudsar, Iran (e-mail: abedini.ac@gmail.com).

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

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, that can be helpful in easily extracting semantic of entities. In addition, since these entities are available in YAGO, so they can be matched with YAGO facts easily.

The contributions of this paper are as follows:

- *Introducing a new ontological fact extraction method called OFE.* In previous approaches, ontological extraction systems tried to obtain new facts that there are not in ontology, but in this paper, to extract ontological facts, the available facts in ontology will be extracted.
- *Introducing a new method to solve semantic relatedness.* There are many methods to compute semantic relatedness problem. But none of them have not used ontological facts. This new method is introduced here.
- *Creating a new application for YAGO ontology*. In this paper using YAGO as background knowledge and facts resource is proposed that can be one of applications of YAGO ontology as a new knowledgebase that has been introduced recently.
- Converting an unstructured text into a set of ontological facts. The method of this paper converts an unstructured text into a set of ontological facts that are available in YAGO ontology.
- *Creating a new application for semantic entity extraction*. Semantic entity extraction method was introduced in author previous work. The ontological facts extraction system that is presented here uses this method to extract ontological facts.
- *Using YAGO relations to optimize ontological facts extraction system.* In final it will be suggested that using YAGO relations such as Type relation can optimize ontological facts extraction system.

This paper has been structured as follows. In next section first our solution for ontological facts extraction is described and then by using it, experimental results will be presented. These experimental results are performed on a benchmark dataset, introduced by Lee [9], and is compared with SOFIE experimental results, as state-of-the-art of ontological facts extraction system, and with ESA [7], as state-of-the-art of computing semantic relatedness. Finally, conclusions are represented.

II. Ontological Facts Extraction

To extract ontological facts from a text for computing semantic relatedness, first semantic entities from the text that are available in YAGO must be extract, then these entities can be matched with YAGO facts. These two steps have come in continue.

A. Semantic Entities Extraction

This extraction has been already performed by author. To extract the semantic entities that are available in YAGO following steps have been performed.

First, the text has been normalized. This means that since characters, dates and numbers can be a semantic 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. A same work in this field has been done in LEILA [10], and its idea has been used in this paper.

Second, the normalized text has been converted to a set of small strings known as tokens. Here the method of SOFIE [2] is used to do this. In this method, a text is given as input and output is a set of tokens with their types.

Third, each one of the tokens has been converted to a semantic entity that is available in YAGO by a token disambiguation method. As mentioned earlier, YAGO ontology is a knowledgebase with high coverage and precision that has been obtained from Wikipedia and WordNet [4]. In fact, it can be said that it is the most appropriate available knowledge resource in mining meaning domain [8]. It contains more than 2 million entities and 19 million facts about them and has only 99 unique relations. So it can be appropriate background knowledge for our 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”.

So by semantic entities extraction 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 [1], so it can be claimed that the semantic entities have been obtained.

B. Matching Entities with YAGO Facts

YAGO facts have more than 95% confidence. So it can be relied. By converting a text to a set of YAGO facts, it can be resulted that context of the text has been gained. Therefore, to computing semantic relatedness of texts, it can help contextual comparing. For this aim, semantic entities that obtained from previous step must be matched with YAGO facts. This matching has been come in algorithm1.

Inputs of this algorithm are a set of semantic entities (*se*) from the text that obtained in previous step, and all facts of YAGO ontology that is shown with *o*. each one of the facts in *o* are as a triple. The triple is included two entities and a relation between them. In this algorithm, set of first entities in facts set *o* is shown by *arg1*, set of second entities in facts set *o* is shown by *arg2*, and set of relations in facts set *o* is shown by *relation.* Also number of facts in *o* is shown by *f.*

In this algorithm, all facts about each one of the semantic entities is obtained. These facts are appeared in *facts* as output.

ALGORITHM OFE

```
Input:   Semantic Entities se, YAGO_Ontology o
Output: Ontological Facts facts
f := number of facts in o
arg1[m] := first entity name in row m in o
```

```
relation[m] := Relation name in row m in o
arg2[m] := second entity name in row m in o
FOR  i = 1 TO n
  FOR j = 1 TO n
    FOR k = 1 TO f
      IF se[i] = arg1[k] OR se[j] = arg2[k]
         {
             Facts[m] := arg1,relation,arg2
             m++
            }
 RETURN facts
```

(1)

The ontological facts that were obtained by this algorithm show all world facts about the text that are available in YAGO ontology. Applications of this method will be shown in next section.

III. Experimental Result

A. Implementation

As mentioned in previous sections, to extract ontological facts in this paper YAGO ontology is used. To implement ontological facts extraction system, 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 about 4.7G that is shown in Table I.

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

Facts table in Mysql has been indexed for faster query answering. The algorithm of ontological facts extraction, have been implemented with java codes on this database.

B. Evaluation

The existing facts extraction systems such as SOFIE as state-of-the-art of these systems have not been designed to compute texts semantic relatedness. In this evaluation it will be shown that OFE is benefit to compute semantic relatedness of texts. For this aim, OFE is performed on a benchmark dataset, introduced by Lee [9], and is compared with SOFIE experimental results, as state-of-the-art of ontological facts extraction system, and with ESA [7], as state-of-the-art of computing semantic relatedness.

OFE has been performed on Lee dataset and frequency of relations of obtained facts is shown in Table II.

Two relations of Means and Type are most relations that extract by OFE. Domain and range of these relations have been shown in table2 that can be helpful to compute semantic relatedness of texts. To clear this topic, a case study is checked for OFE system compared with SOFIE on a text that is shown in Table III.

TABLE II: 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

TABLE III: A Case Study

Text
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.

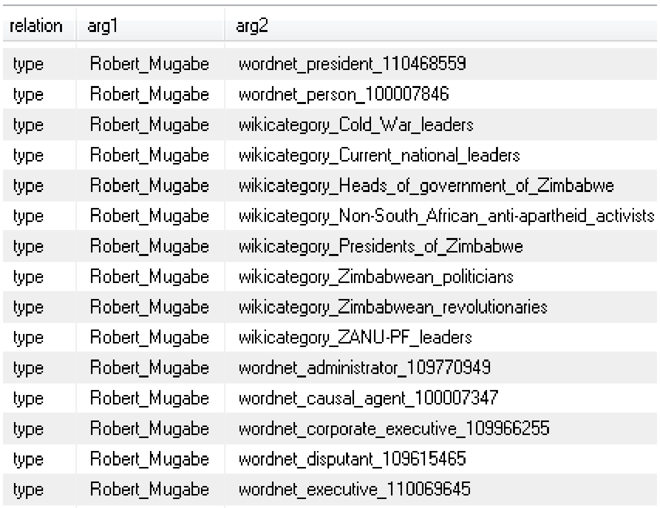
Facts for entities of text by OFE				
Entity	*#Facts*	*#TYPE Relations*	*#MEANS Relations*	*#Other Relations*
United_States	967	16	117	834
Robert_Mugabe	70	25	11	34
Zimbabwe	174	21	10	143
Walter_H._Kansteiner,_III	23	13	6	4
Africa	154	13	11	130

Facts by SOFIE			
Relation	*Arg1*	*Arg2*	*Fact*
disambiguatedAs	^Bush1	Bush_(1916_automobile)	True
disambiguatedAs	^Bush1	Bush_family	false

SOFIE is closest work in this field that uses both of the YAGO and fact extraction from unstructured texts. SOFIE only extract new facts that there are no in YAGO, but OFE extract the facts that available in YAGO.

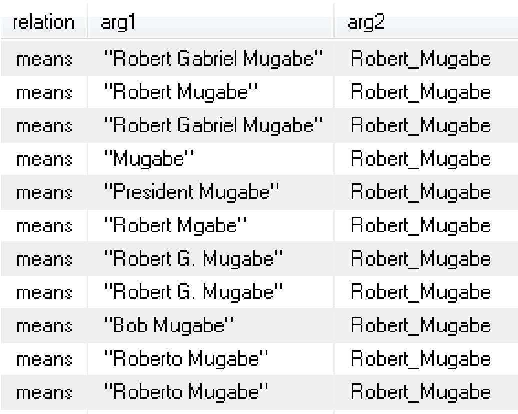
SOFIE facts is little that are not benefit to compute semantic relatedness of texts, but as is clear in Table III number of OFE facts is very much that help us to this purpose. We know that OFE is general and it must be optimized for this application, but such a system able to be used for other applications, because this system has been obtained a set of general facts of the text. Studies have shown that although the Means relation is most relation in extracted facts, but the Type relation is more benefit to compute semantic relatedness of texts. This issue is shown in figure1 and Fig. 2.

In these figures, two relations Type and Means have been shown for the entity of Robert Mugabe from the case study text. In these figures, it is clear that the Type relation is more structured and more useful than Means. In Table II 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 II, upper context of entity will be obtained that help us to compute semantic relatedness easily.



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_Zimbabwe
type	Robert_Mugabe	wikicategory_Non-South_African_anti-apartheid_activists
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

Fig. 1. 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

Fig. 2. Facts of MEANS relaition in YAGO

To show quality of OFE, its entities will be compared with ESA entities in a text in table4. Comparing with ESA, it can be said that most of the extracted words using ESA is not available in related text. These words are name of articles in Wikipedia such that some of words in the text are available in context of those articles. These article names are used to compute texts semantic relatedness. But it has been proposed that entities of OFE can be used for computing semantic relatedness. Power of this suggestion has been shown in following.

The method presented here can obtain ESA entities only by one of YAGO facts relations called FOUNDIN. So, this approach can be more complete than the previous ones. The advantage of OFE is that, for each one of entities all facts of them are available by which semantic relatedness of texts is computed easily. OFE is a very general approach and we are going to optimize it in our next work.

TABLE IV: COMPARITION OF OFE AND ESA

Text	*U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials.*
OFE entities	Saddam_Hussein White_House Central_Intelligence_Agency Iraq Bush
ESA entities	Iraq disarmament crisis Yellowcake forgery Senate Report of Pre-war Intelligence on Iraq Iraq and weapons of mass destruction Iraq Survey Group Iraq War Scott Ritter Iraq War- Rationale Operation Desert Fox

IV. CONCLUSION

In this paper, the approach of extracting ontological facts from a text using YAGO ontology was presented. In evaluation it was shown that our method is benefit to compute semantic relatedness of texts.

In previous approaches, ontological extraction systems tried to obtain new facts that there are not in ontology, but in this paper, to extract ontological facts, the available facts in ontology are extracted.

The contributions of this paper was introducing a new ontological fact extraction method called OFE, introducing a new method to solve semantic relatedness, creating a new application for YAGO ontology, converting an unstructured text into a set of ontological facts, creating a new application for semantic entity extraction, and using YAGO relations to optimize ontological facts extraction system.

Since, the method of this paper extracts ontological facts, and ontological facts are facts about real world, so we can use them to solve open problems such as semantic relatedness. The method introduced here can improve computing semantic relatedness. For this aim, in our next works we are going to use this method to compute semantic relatedness of texts.

In our next work, we consider using the Type and the subClassOf relations of YAGO to optimize OFE to compute semantic relatedness of texts.

REFERENCES

[1] 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.

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

[3] Alexander Budanitsky and Graeme Hirst. *Evaluating wordnet-based measures of lexical semantic relatedness*. Computational Linguistics, 32(1):13–47, 2006.

[4] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[5] Peter Roget. *Roget’s Thesaurus of English Words and Phrases*. Longman Group Ltd., 1852.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. JASIS, 41(6):391–407, 1990.

[7] Gabrilovich, E. and Markovich, S. *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, 2007.

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

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

[10] 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.

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.

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.