Named Entity Corpus Construction using Wikipedia and DBpedia Ontology

Younggyun Hahm[1], Jungyeul Park[2], Kyungtae Lim[3], Youngsik Kim[3]
Dosam Hwang[4], Key-Sun Choi[1,3]

[1]Division of Web Science and Technology, KAIST, Republic of Korea
[2]UMR 6074 IRISA, Université de Rennes 1, Lannion, France
[3]Department of Computer Science, KAIST, Republic of Korea
[4]Department of Computer Science, Yeungnam University, Republic of Korea
[1,3]{hahmyg, kyungtaelim, twilight, kschoi}@kaist.ac.kr, [2]jungyeul.park@univ-rennes1.fr, [4]dshwang@yu.ac.kr

Abstract

In this paper, we propose a novel method to automatically build a named entity corpus based on the DBpedia ontology. Since most of named entity recognition systems require time and effort consuming annotation tasks as training data. Work on NER has thus for been limited on certain languages like English that are resource-abundant in general. As an alternative, we suggest that the NE corpus generated by our proposed method, can be used as training data. Our approach introduces Wikipedia as a raw text and uses the DBpedia data set for named entity disambiguation. Our method is language-independent and easy to be applied to many different languages where Wikipedia and DBpedia are provided. Throughout the paper, we demonstrate that our NE corpus is of comparable quality even to the manually annotated NE corpus.

Keywords: Corpus, Named Entity Recognition, Linked Data

1. Introduction

Named Entity Recognition (NER) is based on a machine learning approach to identify and classify named entities in unstructured text (Nadeau and Sekine, 2007). Most recent NER systems require expensive annotations for training data, called a *gold standard*. There are some available NE gold standard corpora, such as CoNLL-03 (Tjong Kim Sang and De Meulder, 2003), MUC-6 and MUC-7 (Grishman and Sundheim, 1996) provided through their Shared Task, and BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005). However, certain languages have limited research scope where no NE corpus is available. As an alternative to overcome this problem, the automatic corpus construction approach called the *silver standard* corpus has come to the forefront in the field of NE research (An et al., 2003; Nothman et al., 2012)

In the silver standard corpus, NE annotation for NE candidate terms is accomplished automatically, with entity classification and disambiguation tasks. Simple approaches, such as gazettee matching, have limitations because of hyponym or contextual meaning issues (Nothman et al., 2012; Toral and Muoz, 2006).

This paper proposes the method to construct NE corpus using Wikipedia as a raw corpus and SPARQL queries on the DBpedia ontology[1] as a way to classify NEs into DBpedia ontology classes in Section 2 and 3. We show this silver standard corpus has comparable quality against an existing gold standard corpus in Section 4. Especially, our approach is language independent and does not require deeply analysed linguistic features. Hence, in Section 4.3, we show the result of corpus construction for a language that has no available NE corpus.

[1]`http://wiki.dbpedia.org/Ontology`

2. Named Entity Corpus

2.1. Gold Standard Corpus

There are several approaches using linguistic resources for NER modeling, such as WordNet (Toral and Muoz, 2006), extra information in Wikipedia (Cucerzan, 2007; Finkel et al., 2005), and manually annotated corpus (Tjong Kim Sang and De Meulder, 2003). The NE corpus in CoNLL-03 shared task is a main gold standard including NE features. Data from the shared task provide four columns for each word; token, part-of-speech tag, chunk tag, and its NE tag. Here is an example of the CoNLL style format[2]:

```
When        WRB  I-PP  O
Page        NNP  I-NP  I-PER
played      VBD  I-VP  O
Kashmir     NP   I-NP  O
at          IN   I-PP  O
Knewbworth  NNP  I-NP  I-LOC
,           ,    O     O
```

Most NER systems borrow their native format from the CoNLL style corpus, such as the Stanford NER tagger[3]. This is because the CoNLL style format provides segmented, entity classified and disambiguated sentences in structured format. In this respect, a silver standard should provide NE annotated sentences with comparable quality against manual annotation.

2.2. Silver Standard Corpus

In order to construct a silver standard corpus, two tasks should be automatically accomplished: (1) segmentation of

[2]This example was cited in the example of AIDA: `http://www.mpi-inf.mpg.de/yago-naga/aida`

[3]`http://nlp.stanford.edu/software/CRF-NER.shtml`

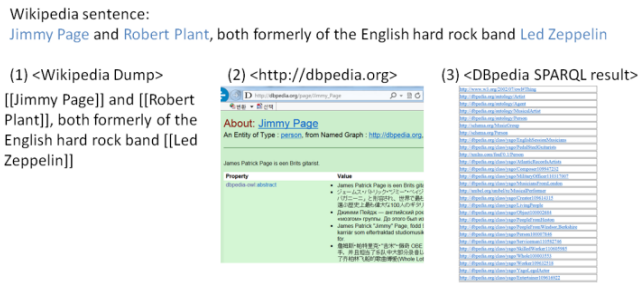

Figure 1: Example of Wikipedia and DBpedia resource for a Wikipedia sentence

the corpus, and (2) NE classification and disambiguation. (Nothman et al., 2012) classify Wikipedia articles and NE tagging for words which link to that articles in their classification granularity. This idea has two advantages: it is easy to obtain a large corpus freely, and it is also language independent. However, article classification would require additional efforts for other language.

In (Rizzo et al., 2012), DBpedia ontology is used as fine-grained NE domains for advantages of accessing and extracting information in the Linked Data Cloud.

Our proposed method follows this Wikipedia approach in construction of the NE silver standard corpus, and it also follows the approach that uses the DBpedia data set and the DBpedia ontology for NE classification. This approach does not require deep analysing linguistic features in Wikipedia, and the resulting NE corpus can also be used for the DBpedia ontology granularity NER system as training data.

2.3. Entity Disambiguation in DBpedia

DBpedia is a dataset that consists of entities extracted from each Wikipedia article (Auer et al., 2007). Each entity can be classified into DBpedia ontology by mapping between Wikipedia article's Infobox and DBpedia ontology's classes (Bizer et al., 2009).

Linked terms in Wikipedia sentences are directed to their Wikipedia target article, and each target article is each entity in DBpedia. Figure 1 shows information about Wikipedia sentences and their linked terms in the Wiki syntax.

DBpedia is a RDF data set, consisting triple S (*subject*), P (*predicate*), and O (*object*). The following triplet shows the result that the types of *Jimmy Page* from SPARQL query.

```
S: <http://dbpedia.org/resource/Jimmy_Page>
P: <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
O: <http://www.w3.org/2002/07/owl#Thing>
<http://dbpedia.org/ontology/Artist>
<http://dbpedia.org/ontology/Agent>
<http://dbpedia.org/ontology/Person>
```

In this case, the DBpedia entity *Jimmy Page* is classified into the DBpedia ontology classes such as Artist, Agent, and Person. Hence, we can use the following SPARQL query to classify linked terms:

```
select distinct ?o where {
<http://dbpedia.org/resource/Jimmy_Page>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
?o }
```

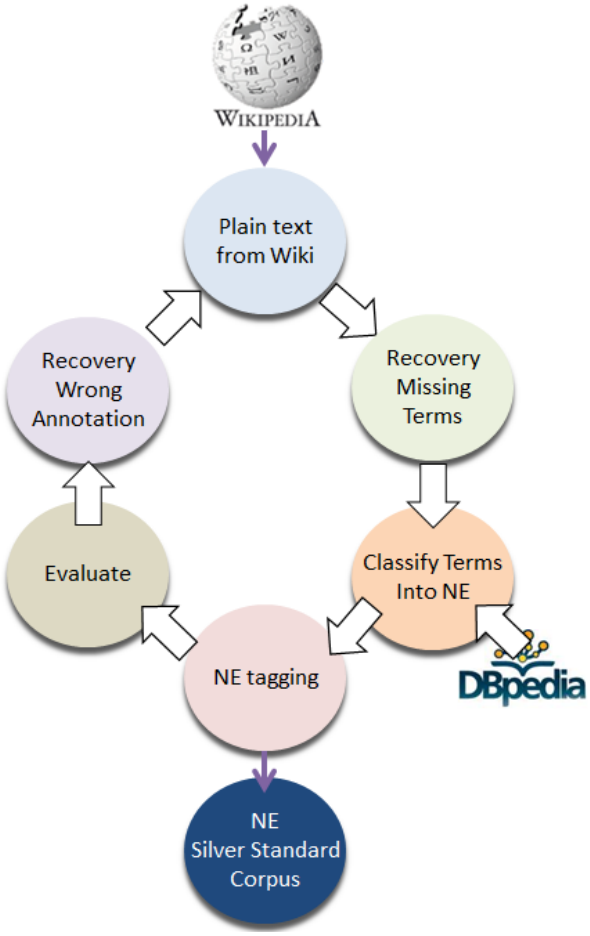

Figure 2: Work cycle for the silver standard corpus construction

In order to construct the NE silver standard corpus, we used: (1) Wikipedia's raw data as segmented corpus, (2) linked terms in Wikipedia as NE candidates, and (3) SPARQL queries for NE classification and disambiguation.

3. Silver Standard Corpus Construction

To construct the NE silver standard corpus from Wikipedia, each linked term should be tagged as NE. This can be presented by the following formula:

$$P_{NE|T} = P(NE|T)$$

where NE is a named entity and T is a named entity candidate term. We assume that T is a linked term in Wikipedia. Additionally, $NEtag$ can be classified by DBpedia SPARQL queries. The work flow is summarized in Figure 2. In this flow, the NE silver standard corpus could show high performance and quality by repeating the work cycle several times.

3.1. Wikipedia Raw Corpus

The Wikipedia dump[4] is an XML format file and includes unnecessary information for our purposes. Hence, we parse plain text with wiki-link annotations from the Wikipedia

[4] http://dumps.wikimedia.org

dump file[5]. At this point, all linked terms are not annotated wiki-link because this task heavily depends on the article writer's efforts, which is not mandatory. The following words are re-tagged to supplement link information in Wikipedia sentences.

1. Surfaceform of Wikipedia article title but not wiki-link tagged
2. Surfaceform of wiki-link tagged word which 1 more tagged in same article

The second task may give rise to hyponym-related problems, but we assume that terms do not require word sense disambiguation (Gale et al., 1992).

3.2. Listing Named Entity from DBpedia

DBpedia ontology 3.9 version has 529 hierarchical classes[6] including typical NER domain PLO (person, location, organization). Our approach can be applied for DBpedia ontology granularity NE classification, but in this paper, we consider PLO domain only to compare the silver standard and CoNLL gold standard. This can be described by the following formula:

$$P_{NE|D} = P(NE|D)$$

where NE is {P, L, O} and D is a DBpedia instance. DBpedia contains 4,004,478 entities. Only 3,255,435 entities are mapped with the DBpedia ontology (Person: 1,124,424 Organization: 329,523 and Location: 755,469). For DBpedia ontology based disambiguation to linked term, we list up NEs. The SPARQL query is as follows:

```
select distinct ?s where
{ ?s
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Person>
 }
```

We can also make lists of NEs' synonym by the following SPARQL query:

```
select * where {
?s <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
 <http://dbpedia.org/ontology/Place>.
?o <http://dbpedia.org/ontology/wikiPageRedirects> ?s
}
```

3.3. Named Entity Candidate Term Tagging

In this section, we show the workflow to construct the NE corpus by matching the Wikipedia raw corpus and the DBpedia entity:

$$P(NE|D) * P(D|T)$$

where NE is a named entity, D is a DBpedia instance, and T is a linked term in Wikipedia.
Using this algorithm, we can annotate all DBpedia ontology classes with the NE tag. In other words, the result S can consist of training data for NER system based on the DBpedia ontology.

[5] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

[6] http://wiki.dbpedia.org/Ontology39

Input: t is a linked term list, n_i is an NE list. A is a Wikipedia article, and i is DBpedia ontology classes, such as PLO
BEGIN
initialize an empty list L_i, S, t
if t *in* n_i **then**
  put t in L_i
end
for *each* L_i **do**
  if *entry of* L_i *is in* A **then**
    extract sentences from A;
    annotate i as a linked term;
    put sentences in S;
  end
end
END

Algorithm 1: ExtractSentences (t, ni, A)

4. Result and Data Analysis

4.1. Silver Standard Corpus

Our NE silver standard corpus construction is carried out on the English Wikipedia dump (Approximately 9.5GB)[7]. We refer such a result as *KAIST silver standard corpus for English*.
Table 1 presents the duration of each step of the corpus construction. In Step 1, we extract plain text from the Wikipedia dump. Step 2 recovers missing terms and list linked terms in plain text. Table 2 shows the number of linked terms in the Wikipedia dump. We recover missing linked terms in section 3.1.

Step	Duration	Result File Size
1: Wiki → plain	20,003 sec	8 GB
2: Linked term listing	429 sec	11 GB
3: NE tagging	1,198 sec	1.3 GB
Total	21,630 sec	1.3 GB

Table 1: Duration of Corpus Construction.

	Before Recovery	After Recovery
Linked Terms	69,419,783	98,794,485

Table 2: Number of Linked Terms in Wikipedia Dump.

4.2. CoNLL gold standard corpus vs KAIST silver standard corpus

Our evaluation methods are based on (Balasuriya et al., 2009)'s approach. Table 3 describes the comparison between the CoNLL-03 corpus and Wiki-DBpedia derived KAIST silver standard corpus. This corpus deals with only PLO NEs to be compared with CoNLL. CoNLL-03 corpus contains 14,987 sentences, 8,215NEs including MISC tags (miscellaneous). KAIST silver standard corpus contains about 6.8 million sentences, about 157 million tokens.

[7] http://dumps.wikimedia.org/enwiki/20130904

Table 4 shows the coverage of the DBpedia NE for CoNLL NE lists. The result presents that LOC has high coverage of 74.68%, but PER and ORG have relatively lower coverage.

	CoNLL-03 (Training set)	KAIST Silver Standard
# of sentences	14,987	6,796,274
# of tokens	203,621	157,396,408
# of NE list	8,215	2,209,416
# of NEs in sentences	23,498	9,522,298
# of NEs per sentence	1.321	1.401
% of NE tokens (PLO)	11.54	6.0498

Table 3: Results of Wiki-DBpedia derived Silver Standard Corpus Construction. In # of NE list, there are 7,345 PLO.

	CoNLL	English DBpedia	
PER	3,613	1,311	36.28%
ORG	2,401	906	37.73%
LOC	1,331	994	74.68%
Total	7,345	3,231	43.99%

Table 4: Coverage for CoNLL.

	KAIST Korean Silver Standard
# of sentences	246,587
# of tokens	7,941,253
# of NE list	35,083
#of NEs in sentences	373,370
# of NEs per sentence	1.5141
% of NE tokens	4.7%

Table 5: Results of KAIST Korean Silver Standard Corpus Construction.

4.3. KAIST Korean Silver Standard Corpus

Our approach can be applied to any other languages, if they can provide Wikipedia, DBpedia, and DBpedia ontology mapping[8]. We also construct the Korean silver standard corpus from the Korean Wikipedia dump[9]. In this case, we count morphemes as tokens. Table 5 shows the result.

5. Conclusion

We automatically construct the NE corpus, which we refer as so-called silver standard using Wikipedia and DBpedia. There are other approaches to construct silver standard but they require additional works to classify terms into NEs. We use Wikipedia sentences, DBpedia ontology and SPARQL queries to classify linked terms in Wikipedia sentences. Our approach has three contributions: First, it is easy to apply to other languages because it does not require deep analysing linguistic feature. Second, it is easy to get a large corpus freely from Wikipedia in about 5.85 hours. Third, the resulting NE corpus can have DBpedia ontology granularity. In other words, an NER system based on our silver standard training corpus, can classify terms into DBpedia ontology classes. It will lead to higher accessibility of NER system to linked data cloud.

As future work, we are planning to extend our method to the multilingual NE training corpus and evaluate them. It can provide better accessibility of the NER system to the linked data cloud. All resource and source code described in the paper are available at the following website: `http://www-nlp.kaist.ac.kr/ner/resource`

[8] `http://mappings.dbpedia.org`

[9] `http://dumps.wikimedia.org/kowiki/20130427`

Acknowledgments

This work was supported by the IT R&D program of MSIP/KEIT. [10044494, WiseKB: Big data based self-evolving knowledge base and reasoning platform]

6. References

Joohui An, Seungwoo Lee, and Gary Geunbae Lee. 2003. Automatic Acquisition of Named Entity Tagged Corpus from World Wide Web. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 165–168, Sapporo, Japan, July. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web, Lecture Notes in Computer Science*, 4825:722–735.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named Entity Recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 10–18, Suntec, Singapore, August. Association for Computational Linguistics.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A Crystallization Point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, September.

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.

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 of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense Per Discourse. In *Proceedings of the Workshop on Speech and Natural Language*,

HLT '91, pages 233–237, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A Brief History. In *Proceedings of COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, August 5-9.

David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1):3–26.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning Multilingual Named Entity Recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Giuseppe Rizzo, Raphaël Troncy, Sebastian Hellmann, and Martin Bruemmer. 2012. NERD meets NIF: Lifting NLP Extraction Results to the Linked Data Cloud. In *Proceedings of the 5th Workshop on Linked Data on the Web*, Lyon, France, 16 April , 2012.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Antonio Toral and Rafael Muoz. 2006. A Proposal to Automatically Build and Maintain Gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Ralph Weischedel and Ada Brunstein. 2005. *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, Philadelphia.