"HmaraNER": A Named Entity Recognizer and Linker

Raabia Asif[1], Mohammad Abdul Qadir[2]

Department of Computer Science,
Capital University of Science & Technology,
Islamabad, Pakistan

[1]raabia.mumtaz@cust.edu.pk, [2]aqadir@cust.edu.pk

Abstract. This paper describes our system, "HmaraNER" for the ESWC Open Knowledge Extraction Challenge 2017 Task 1. HmaraNER is a rule-based system that combines and filters the annotations from a semantic tagger, a named entity recognizer, and a part of speech tagger to produce improved results for the recognition and knowledge base linking of named Person/Place/Organization entities in an English text. For the challenge training dataset, HmaraNER scores F-measure 0.95 for entity recognition task and F-measure 0.78 for entity linking task.

Keywords: OKE Challenge, Named Entity Recognition, Named Entity Linking

1 Introduction

The Open Knowledge Extraction Challenge, Task 1 comprises two subtasks: the recognition of named instances of three classes: Person, Place and Organization in a given text, and the disambiguation of these entities to the DBpedia knowledge base.

2 The System HmaraNER

The system first gets the input text annotated using semantic annotator DBpedia Spotlight [1], the Stanford Named Entity Recognizer [3] and the Stanford Part of Speech tagger [4,5]. The entities identified by Spotlight and NER are first pre-processed, our Merge and Filter algorithm then selects and refines these entities based on certain rules, and assistance from POS tags. Our Query algorithm then queries the identified entities on DBpedia for linking purpose. Each module of the system is now described:

2.1 Pre-processing

Any entities identified by the annotating systems (NER, DBpedia Spotlight) that had no word starting with a capital letter were deleted, as we are only concerned with named entities for this task, which always start with a Capital letter.

2.2 Merge and Filter Algorithm

We define PE, the list of potential entities as:

$$PE = N \cup S \quad (1)$$

N is the list of Capital lettered Entities identified by NER. Each of the 60 sentences in the training dataset is annotated using Stanford NER 3-class and 7-class models and using the online NER demo at corenlp.run [2], the union of Person/Place/Organization identified entities from all three, minus any entity starting with a small letter, makes our list N. Entities identified as City/Country/State or Province are considered Places. In case of clash between NERs' taggings, we give highest preference to corenlp.run's NER, then to NER 7-class and then to NER 3-class. In case an entity recognized by one NER is recognized as multiple entities by other NER, we consider them multiple entities.

S is the set of Capital lettered Entities identified by Spotlight. Each of training dataset's sentences is annotated using DBpedia Spotlight for classes Person/Place/Organization. From the identified entities, those that have no word starting with a capital letter are deleted, the rest making our list S. Since weak annotation is required by the challenge, we consider overlapping entities identified by different systems to be same and add the one identified by NER to the list PE.

We construct a new list E of entities recognized by HmaraNER, from PE by applying the following rules in order.

Rule 1. If an entity in PE is identified to be in other than Person/Place/Organization by NER, this entity is cannot be person/place/organization and therefore is **not** to be added to the list E of entities.

Rule 2. If part-of-speech of a single word entity in PE is identified to be verb, determinant, pronoun, conjunction or adjective, or a multi word entity starts with a conjunction, this entity cannot be a named entity and therefore is not to be added to the list E of entities.

Rule 3. If an entity in PE is identified by both NER and Spotlight, add it to list E.

Rule 4. We observed, many times acronyms were not identified by Spotlight or any NER. A rule was thus formulated to identify acronyms. If an entity in E is immediately followed by parenthesis and inside parenthesis is an acronym for the entity i.e. if these letters are initials of capital lettered words of the entity, then that acronym is also added to the list E and will be linked to the same DBpedia resource as this entity.

Rule 5. An entity identified as Title by NER, if the text immediately following it or immediately followed by it, separated by space or comma space, is a capital lettered proper noun and if this text is not already identified as other than Per-

son/Place/Organization by NER, then we consider this text a Person/Place/Organization entity and add this to our list E.

Rule 6. The NER recognized entities in PE which are not yet decided by any of the above rules to be a part of or not part of E, are added to list E.

Rule 7. If an entity *e* from a text is added to list E, and *e* appears again in the same text, the entity is again added to E. If *e* was identified as person by NER, then even if just first or just second name re-appears in text, this second mention is also to be added to the list E.

Rule 8. For the remaining entities in PE, i.e. the ones that are neither added to E yet and are nor identified to be "not to be added in E" up till now, we check if they have been identified by Spotlight. If yes, and if a DBpedia resource is found for them in linking phase, we add them to list E.

Rule 9. From the entities that were not to be part of E according to Rule 1, if an entity is identified by Spotlight, and only part of it was classed in other than Person/Place/Organization by NER and it contains at least one noun word, then add this entity to E too.

Rule 10. If there are more than one consecutive Place entities in list E which are only comma space separated in input text, and they are on same level in the POS constituency parse tree, we merge them into a single entity and the part before first comma is used for its DBpedia linking.

Rule 11. Entities in PE that were identified to be in class "Nationality" by http://corenlp.run/ NER, and are not yet added to list E, are dealt differently. Instead of just considering the nationality word as a potential entity, we expand this entity to the nouns immediately after it in the same noun phrase, or if it is immediately followed by adjectives and then noun, we expand it to include the adjectives and the noun in the same noun phrase, if the following noun is not already recognized as Person/Place/Organization entity by NER. Entities obtained through this rule will be kept in the list E if a DBpedia resource is found for them in the Entity Linking phase, otherwise these will be removed from list E.

Applying above rules we get list E which will be linked to DBpedia in next phase.

2.3 Entity Linking

For linking purpose, we query the entities in E on DBpedia according to their NER identified type.

Queries.
For all entities in our dataset that are typed person, place or organization, we perform the following DBpedia query with <TYPE> replaced by union of foaf:Person, yago:Person100007846, dbo:Person and schema:Person for person, union of dbo:Place, yago:Location100027167, umbel-rc:Place, yago:Area102735688 and yago:Building102913152 for place and union of foaf:Organization, dbo:Organisation, yago:Magazine106595351 and yago:Group100031264 for organization entities, to extract the DBpedia resource whose label matches the entity label in our dataset. For person entities, if full name of person is there in the sentence, we use full name for DBpedia query.

```
SELECT Distinct ?s WHERE {?s ?p ?o.
?s rdf:type <TYPE>. ?s rdfs:label "<Label>"@en.}
```

Entities for which no result is returned, we check if another entity tag was given by some other NER, if yes we repeat the above query for that label. If not, considering the possibility that the entity could have been incorrectly typed by NER, we perform a relatively general query on DBpedia instead, where we match the entity label in any of the three types, person, place or organization. If still no results are returned, we run a regex query. This query extracts all DBpedia resources whose label **contains** the entity label words in our database. In case we get multiple results, or a disambiguation page as result, the following disambiguation rules are applied in order, till the entity is disambiguated.

Rule 1. Filter from the list of candidates those having type Person/Place/Organization.
Rule 2. Candidate whose label contains most words from the text's entities is considered right.
Rule 3. If the entity had a title (from corenlp.run's NER) adjacent to it in the sentence, the candidate having that title mentioned in its abstract is the right one.
Rule 4. Candidate whose abstract has most entities in common with sentence is right.
Rule 5. If Wikidict entity for this entity was given by corenlp.run and it has at least one word in common with entity label, we link the entity to that DBpedia resource.
Rule 6. The candidate having most common subjects/external links with already linked entities of the text is the right one.

If we could still not disambiguate or if we have no results, we say that DBpedia resource for this entity does not exist and generate a URI for this entity.

Query for Acronyms. Following query is performed for entities that contain all capital letters, after removing any possible periods that the entity label might contain.

```
SELECT Distinct ?s WHERE {?s ?p ?o. ?s rdf:type <TYPE>.
      dbr:<EntityLabel> dbo:wikiPageRedirects ?s.}
```

The entities which were not identified by NER, do not have a class defined. We perform a relatively general query for such entities on DBpedia, where we match entity's label in any of three types. If the whole entity label is a substring of another

entity label, we consider them the same, and DBpedia resource found for one is used to link both. For DBpedia query in such cases, we use the longest string first, and in case no result is returned, we move to shorter substrings for query.

3 Results on Training Dataset

The results of named entity recognition and linking of HmaraNER on the training dataset provided for challenge are given in Table 1.

Total Entities in training dataset	373	HmaraNER Correct Identifications	365
Entities identified by HmaraNER	390	Correct URIs by HmaraNER	299
HmaraNER Identification precision	0.936	HmaraNER URIs precision	0.767
HmaraNER Identification recall	0.968	HmaraNER URIs recall	0.793
HmaraNER Identification F-measure	0.952	HmaraNER URIs F-measure	0.780

Table 1. HmaraNER results on trainig dataset

4 Future Work

We intend to further improve our system in future. New features need to be added when an *'s* is encountered in an entity. Also the DBpedia disambiguation for multiple candidates has much room for improvement that we plan to work on.

References

1. Daiber, Joachim et al. "Improving Efficiency And Accuracy In Multilingual Entity Extraction". *Isem2013daiber*. 2013. Print.
2. Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.
3. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf
4. Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.
5. Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.