An Extensible Schema for Building Large Weakly-Labeled Semantic Corpora

S. Matthew English
The University of Hong Kong
h1395010@connect.hku.hk

Abstract

In NLP data drives research, as evidenced by the frequency with which seminal works of database engineering such as The Penn Treebank have been employed as a basis for experimentation. Traditionally large-scale expertly annotated corpora are expensive and time consuming to produce. This paradigm drove researchers to adopt automated methods for generating labelled data with available tools such as Freebase, DBpedia, and the "infoboxes" found on Wikipedia pages. These knowledge bases have been, or are in the process of being, subsumed by Wikidata, an initiative to concentrate such disparate data repositories in an organized machine readable format. This resource is an important research tool. In this paper, we review our experience using Wikidata in constructing a large annotated corpus under distant supervision, moreover we make the materials, the code used to generate our annotations, freely available to all interested parties.

1 Introduction

Wikidata is potentially a tool of great utility for researchers in computational linguistics. In this work we have attempted to demonstrate a series of mechanisms for deriving value from this resource, to disseminate this information in a structured, immediately accessible format to the NLP community, such that the pipeline we propose herein might serve as the conduit for further investigations.

Distant supervision has emerged as a fundamental component of training and assessment in relation extraction systems. Since it's inception, the data set generated by (Riedel et al., 2010) modeled in part on the work of (Mintz et al., 2009) has been used as the metric by which a large swath of the related literature have benched-marked their relative achievements, including (Surdeanu et al., 2012), (Hoffmann et al., 2011), & (Riedel et al., 2013).

The Riedel dataset uses a factor graph to model the correlation between two entities appearing in the same sentence and is trained by means of constraint-driven semi-supervision. The sentences were drawn from the New York Times (NYT) corpus (Sandhaus, 2008) and the knowledge base (KB) used as the external supervision source was Freebase.

With respect to the distant supervision annotations constructed using Freebase as the remote KB (Surdeanu et al., 2012) makes an important remark regarding evaluation, namely *"(Riedel et al., 2010) observes that evaluating on this corpus underestimates true extraction accuracy because Freebase is incomplete. Thus, some relations extracted during testing will be incorrectly marked as wrong, simply because Freebase has no information on them"*. Here we have it that the circumscribed information maintained by Freebased served as a factor to inhibit the accuracy of those systems.

It was thought that with time Freebase would improve and the scope of information made available through this medium would increase, however that is no longer the case. On March 31^{st} 2015 Freebase became read-only, no longer accepting additions or edits as part of a general process of deprecation, in accordance with the decision *"to help transfer the data in Freebase to Wikidata, and in mid-2015 (June 30^{th}) [to] wind down the Freebase service as a standalone project"*.[1] Accordingly, in this paper we seek to establish a framework within which to carry forth the work of relation extraction systems effected by this development.

[1] plus.google.com/109936836907132434202/posts/bu3z2wVqcQc

2 Related Work

In recognition of the relevance the NYT / Freebase corpus has played in the field of relation extraction there have been attempts to modernize the database with more recent Freebase dumps. The most recent attempt with which we are familiar, the work of (Abad et al., 2014) resulted in a recognized improvement to the results obtained in (Riedel et al., 2010).

We know of no previous work which has investigated the generation of a distantly supervised corpus using Wikidata as a remote KB.

The conception of Wikidata in the semantic web community, most notably the work of (Vrandečić and Krötzsch, 2014) as well as the Wikidata Toolkit API[2] are the foundation upon which it has been possible to conceive of employing the methods and tools described herein toward an extensible schema for building large weakly-labeled annotated corpora.

3 Pre-Processing

The NYT corpus[3] comes formatted in a form of XML which can be regarded as largely superfluous insofar as it adds no evident utility to the data and renders it less human-readable. Accordingly, the first pre-processing measure undertaken was to separate the sentences from the mark-up text using a short script.

Once the mark-up features have been stripped away we are left with textual data partitioned by paragraph as in the following form:

```
<block class="full_text">
<p>An article by Bloomberg News in
Business Day on June 13 about the fraud
trial of Conrad M. Black misspelled
the surname of an attorney for Mark S.
Kipnis, of Montgomery, Alabama.  And a
correction in this space on Friday also
misspelled his surname.  He is Michael
Swartz- not Schwartz or Schwarz.<p>
</block>
```

The relation extraction task has generally been determined on the basis of a scope set by the bounds of one individual sentence within which we consider the mutual correlation between one or more entities, as longer term dependencies can quickly become unwieldy, inducing error. With respect to this restriction we engineered a short script to enhance the functionality of the Stanford CoreNLP (Manning et al., 2014) annotation tool, enabling us to apply its sentence splitting function across our large dataset. The processed output of this operation is then rendered as follows:

```
<sentence> 0
An article by Bloomberg News in Business
Day on June 13 about the fraud trial of
Conrad M. Black misspelled the surname
of an attorney for Mark S. Kipnis , of
Montgomery, Alabama .
</sentence>
```

The procedure for the extraction of relations from text is the determination of a shared attribute between entities in a sentence, and therefore it becomes necessary to recognize the named entities under consideration. For this task we used the Stanford Named Entity Recognizer (Manning et al., 2014), applying it with an adapted script to generate the following output:

```
<sentence>
An article by<ORGANIZATION>Bloomberg
News </ORGANIZATION>in Business
Day on June 13 about the fraud
trial of <PERSON>Conrad M.
Black</PERSON>misspelled the
surname of an attorney for
<PERSON>Mark S. Kipnis</PERSON>,
of <LOCATION>Montgomery, Alabama
</LOCATION>.
</sentence>
```

Our pipeline next employs JSoup[4] to parse the custom XML generated by the Stanford NER tool. In so doing we created three separate files, one for each type of identified entity, which is of consequence in the disambiguation phase of our system. Initially these files contain multiple duplicates, thus necessitating a small addition to our pipeline to ensure they contain only unique values. The final counts amount to:

Number of Unique Entities	
Locations	104,867
Organizations	219,563
Persons	637,124

[2] mediawiki.org/wiki/Wikidata_Toolkit

[3] catalog.ldc.upenn.edu/LDC2008T19

[4] jsoup.org

4 Architecture

4.1 Item By Title

Wikidata is organized so as to facilitate the integration of related information between languages and across pages. This is fundamental to the core system architecture wherein each node is represented by a unique alpha-numeric identifier. That ID in turn is subsequently linked to the alternate aliases by which that entity is known, e.g. formal vs. colloquial appellations or different languages.

For instance the ID *Q5816* is representative of "Mao Tse-tung", "Mao Zedong", "Chairman Mao", "Chairman Mao Zedong" and "毛泽东". All of which can be used as the point of entry in the extraction of relationships between *Q5816* and other entities. As such, the next step in our database generation procedure was to associate so called *"Q values"* with each index in our lists of named entities.

Having collected each of the named entities from the full NYT corpus we retrieved their respective Q values by serving requests with the English name, as it appears in our entity list, to the *"ItemByTitle"* portal of the Wikidata API, the mechanism by which English language representations can be linked to their respective Q values.[5] This raised the issue of disambiguation.

4.2 Disambiguation

A commonly observed phenomena in our corpus is the propensity to begin a paragraph with an explicit mention of an individual, e.g. "Chairman Mao", but subsequently make more familiar references, e.g. "Mao". While such an entity would be recognized by the NER and added to our list it would evoke a "disambiguation page" when presented to the Wikidata ItemByTitle API, along with a list which enumerates the possible entities associated with that moniker.

To resolve this we constructed our pipeline such that ambiguous names would be associated with all the Q values referenced on their disambiguation page, when possible narrowed down to instances of persons, locations, and organizations respectively. For example the English moniker 'Bush' in our corpus is associated with George H. W. Bush, George W. Bush, Jeb Bush, Prescott Bush, and the Bush family among others.

The unique entity lists, once fully processed in this way appear as a JSON object of the following structure:

```
{
  "Hippolyte Charles":["Q10954032"],
  "Hani Durzy":[""],
  "Bush":["Q2743830","Q221997",
          "Q23505","Q324742"],
   ...
}
```

Specific entities typically maintain a single Q value, unknown entities maintain none, and ambiguous entities maintain potentially many. Therefore we have elected to extract relations more thoroughly at the cost of additional computation, this trade-off is considered further by way of example.

The sentence *"Though many in his family had attended Yale University, Bush chose to attend the University of Texas at Austin."* would be correctly labeled as an instance of `educated at( Bush, University of Texas at Austin)` since the property `educated at`, represented in Wikidata as `P26`, is an edge between the node `Q49213`, or `University of Texas at Austin` and the node `Q221997`, `Jeb Bush`. This is not the case for any of the other Q values associated with the index for `Bush`.

5 Relationship Query

Once the data set associating Q values with all entities identified by the NER has been generated we turn to the task of determining the relationship(s) between entities within a sentence. The first step in this process was to download a data dump of all the possible relationship types, known in Wikidata as properties, or *"P values"*, at present there are approximately 1,600.

This component of the program begins by reading in the processed NYT articles from our pipeline, at the tag '**`<sentence> 0`**' it will cache the filename and the number of this sentence. Subsequently the sentence is parsed and the named entities are extracted. At this point the named entities are cross referenced and associated with all of the Q values to which they have been linked previously. We then transmit a request using the Simple Protocol and RDF Query Language (SPARQL) endpoint for Wikidata, the first of its kind, developed by the Center for Semantic Web Research.[6]

Queries of this kind take the following form:

[5] wikidata.org/wiki/Special:ItemByTitle?

[6] ciws.cl

```
PREFIX :
<http://www.wikidata.org/entity/>
SELECT * WHERE {
:Q221997 ?simpleProperty
:Q49213
}
```

When the above query is input to the SPARQL endpoint it will yield the response `http://www.wikidata.org/entity/P69c`.[7] The appended '`c`' is used for simple statement properties in RDF, we can remove the '`c`' and follow the URL to retrieve the property, i.e. the relationship between the two entities. The above query will yield '`educated at`'(`P69`).

6 Data Representation

In generating output data we determined versatility and malleability to be of primary importance. As such the resultant JSON object encapsulates the Wikidata alpha-numeric values, as well as the English language names of those attributes, moreover we store the filename of the article from which the sentence expressing the relation was drawn and the number of the sentence in that document. This representation is demonstrated in Figure 1. The node maintaining the expressed relationship was intended to be most easily accessible to facilitate queries of this nature.

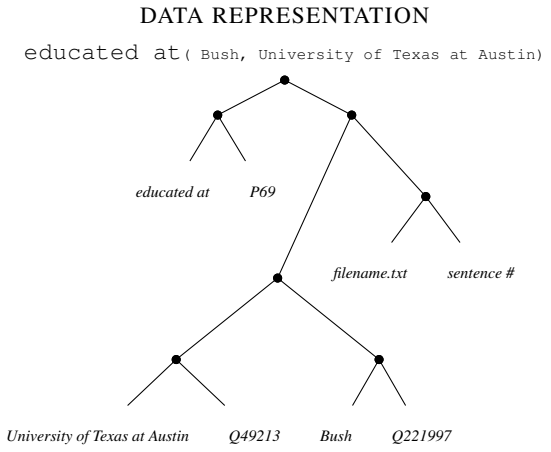
7 Future Work

Relation extraction systems can be susceptible to noise insofar as a sentence that merely contains the two entities that share a relationship in our KB is not necessarily a textual expression of that relationship, consider for instance *"Bush spoke at The University of Texas at Austin commencement ceremonies."* as opposed to *"Bush graduated from UT Austin."*, the former sentence would be misclassified as an assertion of the property '`educated at`'.

Inspired by the work of (Intxaurrondo et al., 2013) our next planned pipeline improvement is to group sentences identified as belonging to a certain class of relation between entities and evaluating them on the basis of Pointwise Mutual Information. Those with a cluster similarity not satisfying an established threshold will be disregarded.

In development of the 'multi-instance multi-label learning' system in (Surdeanu et al., 2012) we are presented with a thorough evaluation of state-of-the-art relation extraction systems on the dataset developed in (Riedel et al., 2010), also used in (Hoffmann et al., 2011). To establish the level of maturity of Wikidata in comparison to its predecessor Freebase we will be replicating these experiments on the NYT / Wikidata weakly-labeled dataset.

Figure 1: An illustration of output datum storage.



8 Conclusion

The software described in this paper are freely available in a publicly accessible code repository.[8]

In this work we have attempted to introduce a simple and extensible mechanism by which the task of creating large-scale distantly supervised corpora can be carried out under present conditions. It is our intention that the tools and procedures we've described will be toward the benefit of the greater community of information extraction researchers going forward.

References

Abad, Azad, and Alessandro Moschitti. 2014. *"Creating a standard for evaluating Distant Supervision for Relation Extraction.". Italian Conference on Computational Linguistics CLiC-it.*, 1.

Intxaurrondo, Ander, Mihai Surdeanu, Oier Lopez de Lacalle, and Eneko Agirre. 2013. *"Removing noisy mentions for distant supervision.". Procesamiento del lenguaje natural 51.*, 41-48.

Hoffmann, Raphael, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. *"Knowledge-based weak supervision for information extraction of overlapping relations.".* Association for Computational Linguistics. *In Proceed-*

[7] milenio.dcc.uchile.cl/sparql

[8] github.com/Building-Large-Annotated-Corpora

ings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. 541-550.

Manning, Christopher D. and Surdeanu, Mihai and Bauer, John and Finkel, Jenny and Bethard, Steven J. and McClosky, David. 2014. *"The Stanford CoreNLP Natural Language Processing Toolkit."*. http://www.aclweb.org/anthology/P/P14/P14-5010 Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 55–60.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. *"Building a large annotated corpus of English: The Penn Treebank."*. Cambridge University Press, Cambridge, UK. *Computational linguistics 19.2*, 313-330.

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. *"Distant supervision for relation extraction without labeled data."*. Association for Computational Linguistics. *In Proceedings of the Joint Conference of the* 47^{th} *Annual Meeting of the ACL and the* 4^{th} *International Joint Conference on Natural Language Processing of the AFNLP.* Volume 2.

Riedel, Sebastian, Limin Yao, and Andrew McCallum. 2010. *"Modeling relations and their mentions without labeled text."*. Springer Berlin Heidelberg. *Machine Learning and Knowledge Discovery in Databases.*, 148-163.

Riedel, Sebastian, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. *"In NAACL-HLT."*. Linguistic Data Consortium, Philadelphia. 74–84.

Sandhaus, Evan. 2008. *"The New York Times Annotated Corpus."*. Linguistic Data Consortium, Philadelphia.

Schoenmackers, Stefan, Oren Etzioni, Daniel S. Weld, and Jesse Davis. 2012. *"Learning first-order horn clauses from web text."*. Association for Computational Linguistics. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* 1088-1098.

Surdeanu, Mihai, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. 2012. *"Multi-instance multi-label learning for relation extraction."*. Association for Computational Linguistics. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* 455-465.

Vrandečić, Denny, and Markus Krötzsch. 2014. *"Wikidata: A Free Collaborative Knowledgebase"*.. Communications of the ACM 57, no. 10. 78-85.

Erxleben, Fredo, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. *"Introducing Wikidata to the Linked Data Web."*. Springer International Publishing In The Semantic Web–ISWC 2014. 50-65.