Datasets and GATE Evaluation Framework for Benchmarking Wikipedia-Based NER Systems

Milan Dojchinovski[1,2] and Tomáš Kliegr[2]

[1] Web Engineering Group
Faculty of Information Technology
Czech Technical University in Prague
milan.dojchinovski@fit.cvut.cz
[2] Department of Information and Knowledge Engineering
Faculty of Informatics and Statistics
University of Economics, Prague, Czech Republic
tomas.kliegr@vse.cz

Abstract. We present a wikifier evaluation framework consisting of software support and two datasets (News and Tweets), which were derived from datasets previously published at WEKEX 2011 and MSM Challenge 2013. Entities recognized in the original datasets were enriched with new annotations – a link to Wikipedia and the most specific type from the DBpedia Ontology. The annotations were created by two annotators and a judge. The datasets are supplemented by plugins for their import to the GATE NLP framework and a DBpedia Ontology-aware plugin for aligning annotations created by a wikifier with the ground truth.

Keywords: Named Entity Recognition and Classification, Benchmark, Wikipedia, DBpedia, Natural Language Processing

1 Introduction

Wikifiers are systems that recognize entities in text and assign them URLs of Wikipedia articles that describe these entities. Some wikifiers also assign types from a taxonomy, making their output comparable with that of Named Entity Recognition (NER) systems. We identify two elementary tasks that the wikifier performs: i) disambiguation (linking of entities to Wikipedia articles, ii) assignment of fine-grained types.

In this paper, we present a framework for evaluation of these two tasks consisting of two datasets, *News* and *Tweets*, and three plugins for the GATE text engineering platform.[1] DBpedia Ontology was selected as the set of types for the fine-grained classification. We have found this choice natural, due to its wide adoption and the fact that a mapping between its classes and the proprietary

[1] http://gate.ac.uk/

types output by many entity classification systems, such as OpenCalais[2], AlchemyAPI[3] and Zemanta[4] is provided by the NERD ontology.[5]

The reminder of the paper is structured as follows. Section 2 reviews available resources for wikifier evaluation. Section 3 describes the contributed News and Tweets datasets. Section 4 presents the evaluation framework. Section 5 covers availability and licenses. Finally, Section 6 provides a concluding summary.

2 Datasets

Currently, there is a lack of resources for wikifier evaluation, since those previously created for benchmarking of NER systems cannot be directly used. While some datasets have been recently contributed, in particular WEKEX[6] or MSM[7], these do not contain all the necessary features for automated wikifier evaluation. The WEKEX dataset contains manual evaluation of results obtained via the NERD framework [2] from multiple common wikifiers. This provides a very useful benchmark of the involved systems at the point in time when the assessment was performed. However, the design of the dataset does not foster its straightforward reuse for a new evaluation. Another recent dataset is the MSM challenge dataset aiming at evaluation of coarse-grained type assignment, and the NIST TAC Entity Linking contest dataset[8], evaluating entity disambiguation. Only the latter provides a comprehensive set of resources for the evaluation of wikifiers (disambiguation only). Unfortunately, this dataset is available for purposes of the TAC contest. None of the listed datasets supports evaluation of fine-grained classification of entities.

In this work, we take up two recently published datasets, the WEKEX dataset and the MSM Challenge 2013 dataset and extend them to fit the needs of Wikipedia-based entity linking and classification, creating the *News* dataset, and the *Tweets* dataset. The two datasets are complementary in that the WEKEX dataset consists of a small number of standard-length news articles, while the MSM datasets contains a large number of very short texts (tweets). Table 1 gives an overview of the size of both datasets.

2.1 WEKEX Dataset

The 2011 paper [2] presents an evaluation of common entity recognition systems using the NERD framework. In this evaluation there were two tasks. We created the *News* dataset from the first task's data. In this task, four participants rated the entities output by the individual systems for ten English news articles

[2] `http://www.opencalais.com/`
[3] `http://www.alchemyapi.com/`
[4] `http://www.zemanta.com/`
[5] `http://nerd.eurecom.fr/ontology/`
[6] `http://nerd.eurecom.fr/ui/evaluation/wekex2011-goldenset.tar.gz`
[7] `http://oak.dcs.shef.ac.uk/msm2013/ie_challenge/`
[8] `http://www.nist.gov/tac/2013/KBP/EntityLinking/`

Table 1. Size metrics for the Tweets and News datasets.

	Documents	Entities			
		All	With CoNLL type	Ontology type	Wikipedia URL
News	10	588	580	367	440
Tweets	1044	1523	1523	1379	1354

selected from the on-line archives of the BBC and The New York Times. These articles were from five different categories. For each entity, a Wikipedia link, if available, and the assigned type, were assessed.

Since each entity recognition tool recognized a slightly different set of entities, the set of all distinct entities identified by the benchmarked systems in [2] were considered for the News dataset.

A limitation of the original WEKEX dataset is a restricted copyright for the underlying textual content. While this textual content is freely available from the BBC's[9] and NYTimes's[10] official websites, it cannot be distributed along with the annotations. The WEKEX dataset is released under the Creative Commons BY-SA 3.0 license. Links to the original content are listed on the dataset website.

2.2 MSM Challenge Dataset

The Making Sense of Microposts (MSM) Challenge 2013 dataset aimed at classifying entities in microposts (tweets). There are four entity types considered corresponding to the standard CoNLL categories: Person, Organization, Location, and Miscellaneous. The dataset is split into two parts – training and test data. Both contained already recognized entities in the text. The entities in the training dataset have the types already assigned, while the types are missing in the test dataset. The organizers also published the goldstandard for the test dataset containing the correct entity types. The original MSM dataset is provided under the Creative Commons BY-NC-SA license.

To construct the *Tweets* dataset, we used the tweets in the goldstandard that contained at least one entity, resulting in 1044 tweets (1523 entities).

3 News and Tweets datasets

The News and Tweets datasets were created by partial reannotation and enrichment of the WEKEX and MSM datasets. The newly created datasets match the needs of automated wikifier evaluation.

[9] `http://www.bbc.com/`
[10] `http://www.nytimes.com/`

3.1 Annotation Guidelines

The WEKEX and MSM datasets were reannotated to a (nearly) common set of fields. The original version of both datasets already provided entity recognition. The annotators thus worked on the same set of entities, providing for each entity:

- **URL to English Wikipedia:** a URL of an article describing the entity,
- **Fine-grained type:** a class from DBpedia Ontology 3.8,
- **Coarse-grained type:** a CoNLL category (only for WEKEX)[11]
- **Most frequent sense flag:** 1 if the correct Wikipedia page is found as the first hit of Wikipedia search for the entity name, otherwise empty.

For the News dataset, we considered as entity candidates each entity output by any of the systems that generated the original annotations in the WEKEX dataset. To deal with this broader scope and lower quality, several specific annotation fields were added to the News dataset:

- **Common entity**: 1 if the entity is not a named entity,
- **Full name**: if this specific entity is a part of a full entity name, which appears in the article, then this fields lists the full entity name,
- **Partial**: 1 if the recognized string is a part of the entity name, and this part does not appear as a full standalone reference to the entity in the document.

The entities for which the result of the annotation process was "not an entity" were removed.

The MSM dataset contains high-quality recognition of entities (with the definition of entity being narrowed to the named entity), therefore entity recognition can be reused in the Tweets dataset. There was just one problem related to entity recognition, which was the frequent incorrect letter casing characteristic for tweets. For the Tweets dataset, there was thus one dataset-specific field added:

- **Incorrect capitalization:** 1 if there is at least one letter in the entity name with incorrect case.

3.2 Annotation process

Each entity was independently annotated by two annotators. If their annotations matched in the specific field (e.g. link to English Wikipedia) this annotation was automatically merged to the ground truth. The annotators were instructed to provide explanation if they felt unsure. When there was no match, another annotator, or in particularly spurious cases two annotators, resolved the conflict, using the explanations from the first-round annotators. The interannotator agreement after the first round (between the two annotators) for the core fields is given in Table 2.

The composition of annotators was as follows: one undergraduate computer science student, two graduate computer science students, one post-doc specializing on ontology alignment (all non-native English speakers). None of the authors took part in the annotation process.

[11] The MSM dataset already contains this information.

Table 2. Interannotator agreement for the core fields. The MSM dataset already contained a coarse grained type assignment, which was not reannotated.

	Wikipedia URL	Coarse grained type	Fine grained type	Most frequent sense
News	0.61	0.65	0.70	0.77
Tweets	0.79	n/a	0.64	0.86

4 GATE Evaluation Framework

To facilitate the use of the newly created Tweets and News datasets, we have developed three plugins for the GATE Text Engineering framework.

The **NewsCorpusBuilderPR** and **TweetsCorpusBuilderPR** plugins load the datasets into GATE. It is assumed that the wikifier being benchmarked is also wrapped as GATE plugin, creating GATE annotations on entities recognized in the documents, with annotation features corresponding to the entity type and Wikipedia URL. We provide a reference implementation of such a plugin for the Targeted Hypernym Discovery[12] entity classification system [1].

Once the wikifier has been run, the correct recognition of entities and of the assignment of Wikipedia URLs can be evaluated by the standard GATE means. We recommend using the GATE Corpus Quality Assurance tool. The evaluation of the assigned DBpedia Ontology Type needs to be performed in an ontology-aware fashion, which is not supported by this GATE tool. Consider the case, when the ground truth type for an entity is `dbpedia:VicePrimeMinister` and the benchmarked tool assigns `dbpedia:Person`. While the string-based comparison performed by the Quality Assurance tool would mark such annotation as incorrect, actually the `dbpedia:Person` is correct with respect to ground truth, albeit more generic.

The developed **OntologyAwareFeatureDiffPR** performs comparison of the assigned types taking into account the hierarchy of the DBpedia Ontology. The plugin also assigns entities a new feature `matchtype`, which has either of the following values:

- exact match: ground truth and wikifier types match,
- supertype: the ground truth annotation is a super-class of the assigned class,
- subtype: the ground truth annotation is a sub-class of the assigned class,
- nomatch: neither of the above.

In case of supertype/subtype match, the plugin also uses the `distance` feature to denote the length of the path between the subtype and supertype classes. Note that the DBpedia Ontology, seen as a taxonomy tree, does not contain cycles and the path length can be thus easily computed.

The plugin also creates a new feature `aligned-type`, which is set to the common supertype (if exists) of the fine-grained type assigned by the wikifier and ground truth fine-grained type. Consider e.g. the following example. For a given

[12] `http://entityclassifier.eu/`

entity, the ground truth annotation contains type `dbpedia:VicePrimeMinister`, while the wikifier assigns type `dbpedia:Person`. The plugin will create the `aligned-type` feature and set it to `dbpedia:Person` on both the ground truth and wikifier annotations. This will allow the native GATE Corpus Quality Assurance tool to evaluate this annotation as correct, while the `matchtype` feature holds the detail type of the match.

5 Availability and License

The reannotated WEKEX (News) and the MSM (Tweets) datasets are available online at `http://entityclassifier.eu/datasets/evaluation/benchmark-datasets/`. The News dataset does not contain the source texts, which need to be obtained from the BBC and NYTimes websites.

Same licenses are used as for the original datasets: Creative Commons BY-SA license for News and Creative Commons BY-NC-SA license for Tweets.

The evaluation framework – the **NewsCorpusBuilderPR**, **TweetsCorpusBuilderPR** and the **OntologyAwareFeatureDiffPR** plugin, together with reference implementation of a GATE PR plugin providing annotations from the `entityclassifier.eu` wikifier, are available online at `http://entityclassifier.eu/datasets/evaluation/tools/`. These plugins are provided under the GNU Lesser General Public License 3.0.

6 Conclusion

While there are several ground truth datasets for evaluation of "classic" NER systems, a freely obtainable dataset for evaluation of Wikipedia-based NER systems does not, to the best of our knowledge, exist. In this paper we proposed an framework and two complementary datasets for benchmarking Wwikifiers, which will hopefully significantly reduce the effort required to perform the evaluation. Sample results of wikifier benchmarks are available on the framework's website. The authors will include links or information on additional results obtained with the News and Tweets datasets if notified.

Acknowledgements. This research was supported by the European Union's 7th Framework Programme via the LinkedTV project (FP7-287911) and CTU in Prague grant (SGS13/100/OHK3/1T/18).

References

1. M. Dojchinovski and T. Kliegr. Entityclassifier.eu: Real-time classification of entities in text with Wikipedia. In H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, (eds.) *Machine Learning and Knowledge Discovery in Databases*, vol. 8190 of *Lecture Notes in Computer Science*, pp. 654–658. Springer Berlin Heidelberg, 2013.
2. G. Rizzo and R. Troncy. NERD: Evaluating named entity recognition tools in the web of data. In *ISWC'11, Workshop on Web Scale Knowledge Extraction (WEKEX'11),October 23-27, 2011, Bonn, Germany*. 2011.