Improving Ontology Service-Driven Entity Disambiguation

A. Patrice SEYED [a] Zachary FRY [b] and Deborah L. MCGUINNESS [b]

[a] *3M HIS, Silver Spring, MD*

[b] *Rensselaer Polytechnic Institute, Department of Computer Science, Tetherless World Constellation, Troy, NY*

Abstract.

One of the long-standing challenges in natural language processing is uniquely identifying entities in text, which when performed accurately and with formal ontologies, supports efforts such as semantic search and question-answering. With the recent proliferation of comprehensive, formalized sources of knowledge (e.g., DBpedia, Freebase, OBO Foundry ontologies) and advancements in supportive Semantic Web technologies and services, leveraging such resources to address the entity disambiguation problem in the industry setting as "off the shelf" within natural language processing pipelines becomes a more viable proposition. In this paper, we evaluate this viability by building and evaluating an entity disambiguation pipeline founded on publicly available ontology services, namely those provided by the NCBO BioPortal. We chose BioPortal due to its current use as an ontology repository and provider of ontological services for the biomedical informatics community. To consider its usage outside the biomedical domain, and given our immediate project goal for facilitating semantic search over Earth science datasets for the DataONE project, we focus on the disambiguation of geographic entities. For this work, we leverage NCBO's Term service in conjunction with NCBO's entity disambiguation service, the Annotator, to demonstrate an enhancement of the Annotator service, through application of a vector space model representation of ontological entities and relationships to drive scoring improvements. This work ultimately provides a methodology and pipeline for improving publicly available ontology service-based entity disambiguation, demonstrated through an enhanced version of the NCBO Annotator service for geographic named entity disambiguation.

Keywords. entity disambiguation, ontology, geospatial

Introduction

One of the long-standing challenges in natural language processing is uniquely identifying entities in text, which when performed accurately and with formal ontologies, supports efforts such as semantic search and question-answering. Semantic search would greatly facilitate our project, the Data Observation Network for Earth (DataONE), as it is aimed at limiting the excessive time and effort spent to discover, acquire, interpret, and use related data for biological, ecological, environmental and Earth science data.[1]

[1] http://dataone.org

The DataONE metadata catalog is composed of content uploaded by participating research institutions, that includes a *keywords* field populated by data managers, where these keywords and keyword-“worthy” terms in the scientific abstracts are not yet linked to domain knowledge in a formal way to aid discovery, besides what exist as disparate controlled vocabularies. Clearly here, the use of publicly available formal ontologies to disambiguate terms of relevance for search provides advanced capabilities for precise search and a “free extension” to their existing vocabularies that does not require manual development.

With the recent proliferation of comprehensive, formalized sources of knowledge (e.g., DBpedia,[2] Freebase,[3] OBO Foundry ontologies[4]) and advancements in supportive Semantic Web technologies and services, leveraging these resources to address the entity disambiguation problem in the industry setting as “off the shelf” within natural language processing pipelines becomes a more viable proposition. In this paper, we evaluate this viability by building and evaluating a novel entity disambiguation pipeline founded on publicly available ontology services, namely those provided by the NCBO BioPortal. We chose BioPortal due to its central role as a ontology repository and provider of ontological services for a given community, that being biomedical informatics. To consider its usage outside the biomedical domain, and given our immediate project goal for facilitating semantic search over Earth science datasets for the DataONE project, we focus on the disambiguation of geographic entities, using the Gazetteer Ontology (GAZ).

In this work, we use NCBO’s Term service in unison with NCBO’s own entity disambiguation service, the NCBO Annotator, to demonstrate an enhancement over the Annotator service, using a vector space representation of ontological entities and relationships to drive scoring improvements. Fittingly then, we evaluate our results against the NCBO Annotator, and apply the TopN scoring method, since both systems are suited to provide inputs to a semi-automated semantic-mapping workbench environment. Ultimately, this work evaluates whether our techniques improve on TopN mapping accuracy of the Annotator service, including if introducing all domain-level relationships into the vector space provides additional discriminatory power over just those relationships considered “broader”, while at the same time provides insights into the process of using and augmenting publicly available ontology-service driven web services for entity disambiguation.

Our overall approach extracts named entity labels from natural language text, and where applicable, maps each to a resource of an ontology that best represents and disambiguates its meaning. The set of entities that can be disambiguated within our pipeline is flexible, to disambiguate what in the formal ontology community is referred to as particulars or concepts (or types), and what in the Web Ontology Language (OWL) considers individuals or classes,[5] that the Semantic Web community at large via the Resource Description Framework (RDF) refers to simply as resources.[6] Thus we describe our pipeline at the abstraction of “resource”, and where appropriate for its current application we describe how it functions for geographic named entities disambiguation. In the following sections we describe related work, background on the topic-based vector

[2] http://dbpedia.org/

[3] http://www.freebase.com/

[4] http://www.obofoundry.org/

[5] http://www.w3.org/TR/owl-ref/

[6] http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

space model algorithm we apply, and our methodology and pipeline that utilizes these algorithms and services. In Section 4 we evaluate results using the TopN scoring metric, comparing against the existing annotation service using an initial, hand-curated gold standard. In Section 5 we present qualitative findings that resulted from application of our the pipeline, and in sections 6 and 7 we discuss future work and conclusions.

1. Background on eTVSM

In this section we formally present the definition and implementation of an eTVSM model [10][5]. An eTVSM formalizes how resources represent named entity labels by first including a TVSM of the ontology, and then representing entity labels within the TVSM based on links between named entity labels and resources. The first step in building an eTVSM is to encode a graph representation of resources connected through a given set of relationships into resource vectors that compose a TVSM. This graph representation is based on candidate resources and their related resources. (We consider candidate resources those resources which are discovered by initial lexical matching and potentially represent to what a named entity label refers.) For each resource, we consider $P(r,k)$ to be the power set of all resources at distance k from resource r. We construct a resource vector $\vec{r}$ as follows:

$$\vec{t} = \left\langle \sum_{k=0}^{\beta} \sum_{r_1 \in P(r,k)} \alpha^k, \ldots, \sum_{k=0}^{\beta} \sum_{r_n \in P(r,k)} \alpha^k \right\rangle, \vec{r} = \frac{1}{||\vec{t}||}\vec{t} \tag{1}$$

Each resource in the graph representation of an ontology is assigned an index from 1 to n, which means that each resource vector has size n and there are n resource vectors. The vector for resource n is calculated by assigning an exponentially declining weight to connected resources and summing the total weights for each resource. For example, resource i is assigned a weight of 1, all resources one node away from i adds a weight of 0.5, resources two nodes away from i add a weight of 0.25, etc. The sum of the weights are normalized. In this way we assign a resource vector to each candidate resource. The constant α is an exponential decay which is used to determine how important resources are distant from the candidate. [5] experiments with different decay values and identify .9 as being optimal for the DBpedia Ontology; in our work we apply an exponential decay constant of .5 and leave this analysis for future work. The constant β is used to limit the distance between resources that we wish to consider; in our case β is 5.

Next, an interpretation vector is constructed from an interpretation, i, which is a unique mapping of a named entity label to a resource, which we multiply by an ambiguity weight:

$$\vec{i} = \vec{r}\, g(i) \tag{2}$$

In order to reduce the effects of ambiguous mappings (i.e., which map labels to multiple resources), we weigh each interpretation vector by an ambiguity weight $g(i)$, defined as:

$$g(i) = \frac{1}{|j : j \in I(k), k \in K(i)|} \tag{3}$$

The ambiguity weight $g(i)$ forces labels which are mapped to many resources to have less weight toward disambiguating other named entity labels. Here, $K(i)$ is the set of all labels for interpretation i, and $I(K)$ is the set of all interpretations derived from label k. We represent the document as the collection of all resource mappings by summing the weighted interpretation vectors:

$$\vec{t}_d = \sum_{i \in I} w_{d,i} \vec{i} \tag{4}$$

Each interpretation vector is weighted by multiplying the frequency at which the named entity label for interpretation i is found in the extracted named entities from document d. The cosine similarity metric is used to compare individual interpretation vectors for the document against the document vector, $\vec{t}_d$. Incorrect interpretations will not significantly affect the document vector and result in a low cosine similarity score, whereas interpretations that are more similar to other interpretations will result in a higher cosine similarity.

2. Pipeline

We apply a three-phase pipeline for executing our entity disambiguation approach, including NER, resource mapping, and eTVSM scoring. We describe each phase as a black-box system, with inputs and outputs which link the phases together. Our overall pipeline takes as input an unstructured text document in the form of a science publication abstract from DataONE outputs a ranked list of candidate resources.

The *NER phase* of our pipeline extracts named entities embedded in unstructured text using the Natural Language Tool Kit (NLTK) [11]. This phase serves the sole purpose of generating a set of labels to be processed for resource mapping in the next phase. The NLTK software library contains NER algorithms that extract named entity labels from unstructured text. It applies a series of tokenizers to parse sentences into terms, and using part-of-speech tagging generates parse trees as input to supervised learning algorithms for named entity detection. The unstructured text which is passed through the NER phase of our pipeline outputs a list of named entity labels for each given document.

The *resource-mapping phase* of our pipeline processes a list of named entity labels and returns a set of named entity label to resource mappings and an ontology portion (i.e., a set of statements) describing each matched resource. In the context of geospatial named entity recognition, there are multiple resources which are referred to by the same preferred label annotation. For example, in the United States, the city of Springfield can refer to greater than 40 different cities located in multiple states. Also some states, such as Wisconsin, have multiple cities named Springfield. Therefore, the purpose of the resource-mapping phase is to map each named entity label to those resources which define a unique instance of that particular location with that name. For every named entity label extracted from the NER phase, we query the NCBO Annotator Service for a list of candidates, that is, resources which could potentially define that particular named entity within an ontology.

Next, we obtain statements about the resources, including annotations as well as object property and subclass statements, using the NCBO Term Service, by supplying a resource URI and ontology identifier as parameters to the service.[7] The object property and subclass-based statements are filtered according to a preselected set of properties, where the objects of these statement are recursively submitted to the Term Service. In this way, we perform a breadth-first search along a set of properties until the final resource in the path is reached. This phase concludes when the statements are collected for each candidate resource.

The *eTVSM scoring phase* scores each entity label-to-resource mapping in a given document by first generating a TVSM of the obtained ontology graph, and then encoding the resource mappings into an eTVSM. We use the cosine similarity metric to determine how strongly one candidate resource is related to all candidate resources for a given document [5]. Given an eTVSM that encodes each mapping of a named entity label to candidate resources into a vector space, closely related candidate resources will have higher scores than those than are more indirectly related or not related at all to other candidate resources. Ultimately, the eTVSM scoring phase provides a quantitative measure for how strongly a resource disambiguates a named entity label.

3. Evaluation

In this section we introduce a small hand-annotated dataset which we use as a gold standard for comparing our result against the existing NCBO Annotator Service. We score output from the NCBO Annotator Service and from our pipeline using the TopN scoring metric. We demonstrate how our approach contributes an enhancement to the NCBO Annotator Service for geographic named entity disambiguation using the eTVSM algorithm. The process for which we created our gold standard is described below.

First, we selected scientific paper abstracts by processing each abstract through our pipeline and selecting those which have at least one named entity label that has greater than 10 candidate concept mappings. This decision was due to a lack of preexisting gold standard data and our desire to disambiguate highly ambiguous named entity labels. For each named entity label we hand-annotated it with a concept in the Gazetteer Ontology which correctly identifies it. The result is 24 unique named entities labels, which were extracted using the NER phase of our pipeline. For these 24 named entity labels, there are 18 unique candidate concepts contained in the GAZ ontology, which we discovered by querying the annotator service with our named entity labels. The remaining six named entity labels were either too generic for us to assign a unique concept, or the correct concepts are not contained in the most current version of GAZ; in order to get a meaningful comparison of our pipeline against the NCBO Annotator Service we excluded these six labels for our evaluation. The 18 named entity labels were hand-mapped to concepts in GAZ, composing our gold standard dataset which we use in our evaluation.

In the setting of a semantic-mapping workbench, when a prospective curator inspects the list of candidate concepts for a set of named entities extracted from a document, the most accurate or correct concept will ideally appear at or near the top of a ranked list of candidate concepts. In fact, the Annotator website presents candidate concepts in this ranked manner. To be fair, there are just two levels of ranking produced by the Annotator

[7] We describe the atoms of the ontology portion extraction in terms of statements, to maintain our abstraction at the level of resources. In the case of named entity disambiguation these statements are a mix of assertions and axioms, and in the case of conceptual entities these statements are primarily axioms.

service, determined by whether a match was made on the preferred or alternate term, while our approach uses a much more granular similarity score metric. Still, we evaluate our augmented version of the annotator alongside the publicly available Annotator Service in order to demonstrate improvement upon the existing service. Therefore we evaluate the two approaches using the position of the correct concept returned in a ranked list of concept-mapping scores. To do this, we use the TopN scoring metric [12,5]:

$$TopN = \frac{\sum_{j=1}^{N} \alpha_c^{k_j}}{\sum_{i=1}^{N} \alpha_c^{i}} \tag{5}$$

where N is the number of named entity labels which are mapped to a concept in the gold standard. k_j is the position of concept j in the returned concepts from the Annotator Service. α_c is an exponential decay coefficient used for penalizing concepts that appear later in the list. A score of 1 is realized for a document if for all named entity labels, the top scoring candidate concept corresponds to the gold standard's concept for that document. Note that since we only consider concepts that are in our gold standard, and by definition, are returned by the Annotator Service, the TopN score can never have a value of 0. However, if all the correct concepts for each entity are returned at the bottom of the list, the TopN score will be significantly small. We chose a value of 0.8 for our constant α_c, so, for example, the position of the correct concept will contribute 0.1 to the score if it appears in the tenth position [5].

For the evaluation of our pipeline, we construct our vector space using two methods. The first construction uses *located_in*, while the second uses all (13) relationships contained in the GAZ ontology. This comparison provides some evidence that including relationships beyond those considered generalizations improves scoring. Table 1[8][9] shows the comparison between the NCBO Annotator Service and our pipeline as constructed using the two methods. These results demonstrate that the application of the eTVSM-scoring phase outperforms the Annotator Service for disambiguating geographic named entities. The results demonstrate that a quantitative approach for disambiguation which measures and ranks the strength of each concept mapping outperforms an approach which relies on simple lexical matching.

4. Qualitative Findings

In this section we present some qualitative insights of applying our pipeline to a sample scientific publication abstract. Figure 1[10] shows a graph output of our pipeline as applied to a sample scientific abstract. In this run we included all geographic relationships from GAZ during the concept-mapping phase.

We confirmed that our ontology-based resource mapping approach successfully down ranks irrelevant resources and thus outperforms purely lexical-based resource map-

[8] A version of this paper with figures included is available at http://tw.rpi.edu/web/doc/ImprovingOntologyServiceDrivenEntityDisambiguation/.

[9] https://www.flickr.com/photos/127739444@N02/15066654769/

[10] https://www.flickr.com/photos/127739444@N02/15230553716/

pings. For instance, where only lexical matching algorithms are used (e..g., the current Annotator Service), the entity label ‘Oregon’ receives equal ranking for “State of Oregon (GAZ:00002515)” as resources “Oregon (GAZ:22225751)” and “Oregon (GAZ:00084619)” (cities in Michigan and Illinois). Our approach leverages fine-grained geographic relationships and considers relationships with other mapped named entities mentioned in the same abstract (e.g., Deer Creek, Josephine County), so that the similarity score of the mismatches for ‘Oregon’ is significantly lower than the correct one.

We also identified improvements to the scoring of correct resources after applying additional relationships in the resource-mapping phase of our pipeline, beyond those considered generalizations (e.g., *located_in*). For example, for the abstract of the fourth study of our gold standard, the named entity label ‘Cumberland River’ was mapped to three distinct candidates. When only the *located_in* relationship is applied, the two incorrect resources received a score while the correct resource (GAZ:00150754) did not (see left side of Table 2[11]); however, when applying all geographic relationships, the correct resource accurately received the highest score. This is due to relationships to other resources, via the inclusion of additional kinds of geographic relationships, that have been mapped to other named entity labels extracted from the abstract (shown in Figure 2[12]).

Finally, we learned that our pipeline improves the quality of ontologies available through Bioportal, by facilitating curators in the process of identifying and reporting existing gaps. For example, in an abstract that mentions the Deer Creek Field Station and Educational Center of the state of Oregon, the named entity label ‘Deer Creek’ returned 29 unique resources labeled as ‘Deer Creek’, none of which were the correct one. We informed the GAZ team, who quickly created the resource and appropriate statements, increasing coverage of the ontology. Figure 1 illustrates the results of the pipeline after the newly added resource “Deer Creek (GAZ:00633440)” was included, which became the highest ranking candidate resource for ‘Deer Creek’ due to relationships to “Josephine County” and “State of Oregon”.

5. Related Work

In this section we discuss recent work that applies ontologies to the named entity recognition (NER) problem, including that which the current work uses and builds upon: the NCBO Annotator and enhanced topic-based vector space modeling (eTVSM). Researchers at the BBC experimented with eTVSMs [10] to automatically apply editor tags to archived radio programs for use in a manual curation environment [5]. Concepts contained in the DBpedia ontology were represented in a topic-based vector space model (TVSM), a model constructed by creating vectors for each concept that include those concepts related by SKOS *broader*[13]. The eTVSM was built by linking text transcribed from radio programs to concepts in the DBpedia Ontology,[14] scoring each link using the relationships between concepts that were encoded in the TVSM. Links that were closely related scored higher, while incorrect links which were not as closely related scored lower. Our work reuses the same underlying theory for using a vector space model for

[11] https://www.flickr.com/photos/127739444@N02/15066727059/

[12] https://www.flickr.com/photos/127739444@N02/15253595325/

[13] http://www.w3.org/2004/02/skos/

[14] http://wiki.dbpedia.org/Ontology

disambiguation, and additionally explores the benefits of using relationships more explicit than *broader*, to take advantage of knowledge beyond that found in a generalization hierarchy, formalized by expert curators.

The NCBO BioPortal project supports efforts to linking unstructured text to ontologies through publicly accessible services for leveraging community based ontologies [2]. The NCBO Annotator Service matches inputted text to ontological terms contained in community-developed ontologies by applying a lexical string matching algorithm to a lexicon based on preferred and synonym labels [3,4].[15] By default, the Annotator Service is configured to consider all ontologies published through BioPortal, however there is a parameter for restricting it to a set of target ontologies. [3] highlights the additional need for enhancing the service by developing components that use the knowledge in ontologies to recognize relationships between concepts, which is a focus of this paper. The service returns a list of candidate concepts from the selected ontologies and provides a score for each candidate concept based on whether the concept was matched on preferred label or synonym. Our approach builds on this service and ultimately creates an enhanced version of it that quantitatively measures how suitable each candidate concept represents a named entity. Further, our approach and pipeline starts by recognizing tokens in the text, while the Annotator spots named entities using terms from the target ontologies and supporting lexicons. Therefore with our approach a curator is more easily able to find and report gaps in the existing ontologies in a semantic-mapping workbench setting, since extracted token are immediately available for inspection. We describe how we practically applied this mechanism in Section 3.

Aside from the BBC and NCBO efforts, there exists extensive previous research in the area of entity disambiguation leveraging ontology or more generally, linked data sources. Alexopoulos et al. [6] propose a disambiguation framework that utilizes DBPedia to detect intended meaning of named entities (e.g., soccer clubs, organizations) in unstructured text, using an algorithm similar to [10]. Kleb et al. [8] focus on disambiguation using spreading activation on an RDFS-based ontologies. Mendes et al. [9] provides disambiguation and mapping to DBpedia URIs, within DBpedia Spotlight. Hoffart [7] applies a novel collective disambiguation strategy using a new form of coherence graph using DBpedia and Yago. There are also many off-the-shelf concept extraction tools available: Open Calais,[16] Zemanta,[17] Alchemy API[18]); all of these approaches identify entities and generate URIs for them through disambiguation.

Our work differs from these in that we focus on the practicality of using NCBO Bioportal and its APIs as an ”off the shelf” resource for applying eTVSM for semantic disambiguation within an NLP pipeline. BioPortal is of particular interest because the ontologies registered with it include many that are developed by expert curators. The benefit of our approach is that the more explicit relationships and to what resources they relate are used for disambiguation and subsequently for fine-grained semantic search capabilities. What results from our work is an enhanced version of the NCBO Annotator for geographic entity disambiguation. Due to the Annotator’s wide usage, it provides immediate utility to the community upon release.

[15] http://bioportal.bioontology.org/annotator

[16] http://www.opencalais.com/

[17] http://www.zemanta.com/

[18] http://www.alchemyapi.com/

6. Future Work

In future work, we will expand our gold standard to help determine if they further validate our results. Since it is a time-intensive process, we will seek external resources for performing the work, such as Mechanical Turk.[19] We will also leverage such resources for tagging corpora for geographic named entities, which can be used for statistically training the NLP tokenizer that we employ in the pipeline.

Our methodology and pipeline lays a foundation for widening its use to conceptual entities and other types of named entities. Therefore, in future work we will evaluate how well, in practice, that our results are generalizable for disambiguating concepts and named entities for other domains, such as biomedical. On the side of named entities, one requirement is to request the NCBO to add additional ontologies that are specific to individuals, similar to the crowdsourced content available via Dbpedia.

For concept-based disambiguation, we lose the immediate benefit of the NLTK named entity recognizer, but which is mitigated when corpora tagging is carried out for the concept domain of interest, via Mechanical Turk or use of some other resources (e.g. PubMed). While there exists a wide range of biomedical ontologies available that cover similar sets of concepts, we will incorporate and test publicly available mappings between NCBO-registered ontologies, though performing the mapping task itself falls out of our scope. For the ontology extraction task of the resource-mapping phase, the mechanism for obtaining axioms at the class level instead of assertions at the instance level remains the same via the NCBO Term service.

In cases where concept mappings are not available, selecting the ontology to use that provides the best coverage and overall representation becomes more critical, as the eTVSM approach requires the selection of one ontology. This selection should be an automated process, as within the context of an annotation software tool for semi-automated mapping, it reduces burden on the annotator, enabling them to focus on finding the most accurate concept match in a ranked list of candidates. Therefore, in future work we will leverage a domain classifier for selecting the most suitable ontology for disambiguation.

To further support annotation software tools that leverage our pipeline, we will make modifications to capture the statements in RDF and/or OWL for ease of rendering in graph form; currently we are applying the XML-based results from the NCBO services in a non-RDF graph representation for processing into the vector models. We anticipate that, generally, the graph representations (as shown in Figure 1), when presented, will provide a curator context and visual justification of the ranking scores. Finally, at the time of this writing the NCBO ontology service for the version of BioPortal being used is deprecated, therefore we are working to port our code to leverage the latest version prior to making it publicly available.

7. Conclusions

To help address the challenge of using publicly available ontology services for entity disambiguation, in this paper we 1) provide an enhanced version of the NCBO Annotator service for geographic named entity through novel application of vector space model-based disambiguation in concert with existing NCBO Term and Annotator services; 2)

[19] https://www.mturk.com/

demonstrate that using the available fine-grained relationships in an expert-curated ontology improves disambiguation; and 3) provide insights into the process of using publicly available ontology-service driven web services and expert-curated domain ontologies for entity disambiguation and organically improving upon those services.

In support of future semantic mapping workbench applications, this approach provides a ranked list of results using quantitative scoring methods to disambiguate named entities. In Section 4 we evaluated the performance of our pipeline against the Annotator Service using the TopN scoring metric and demonstrated how in the context of a manual curation workbench, our pipeline provides benefit by reducing the time a curator would spend looking for the concept that correctly matched an extracted named entity label. To further demonstrate its value as an enhanced version of the NCBO BioPortal Annotator Service, in Section 5 we presented some insights resulting from the pipeline being applied to Earth and environmental science abstracts, leveraging domain-level relationships available in GAZ to power the disambiguation process.

Our approach also helps aid curators to create gold standard datasets as training data for performing entity disambiguation using statistical machine learning methods. For curators who manage metadata like those within the DataONE project, the output from our pipeline could be added as metadata, improving metadata quality by showing how named entities in the text are related, which can we used to enhance search capabilities. Our approach provides benefits over using methods that do not rely on ontologies for the disambiguation task, or when ontologies with minimal semantics (e.g., *broader* relationship in SKOS) are used, as subsequent search interface capabilities will have better precision.

References

[1] Heath, T. and Bizer C. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web. Morgan and Claypool Publishers, 2011.

[2] Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B; NCBO team. The National Center for Biomedical Ontology. J Am Med Inform Assc. 2012 Mar;19(2):190-5. Epub 2011 Nov 10.

[3] Jonquet C, Shah NH, Musen MA. The open biomedical annotator. Summit on Translat Bioinforma. 2009 Mar 1;2009:56-60. PubMed PMID: 21347171; PubMed Central PMCID: PMC3041576.

[4] Shah, N., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A., & Musen, M (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. BMC bioinformatics, 10(Suppl 9), S14.

[5] Raimond, Y., & Lowis, C. Automated interlinking of speech radio archives.

[6] P. Alexopoulos, C. Ruiz, J.M. Gmez-Prez (2012), Scenario-Driven Selection and Exploitation of Semantic Data for Optimal Named Entity Disambiguation, Proceedings of the 1st Semantic Web and Information Extraction Workshop (SWAIE 2012), Galway, Ireland, October 8-12, 2012.

[7] Hoffart, J., Yosef, M.A., Bordino, I., Frstenau, H, Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACL, Stroudsburg, PA, USA, 782-792.

[8] Kleb, J., Abecker, A.: Entity Reference Resolution via Spreading Activation on RDF-Graphs. In Proceedings of the 7th ESWC, pages 152-166, Springer Berlin, Heidelberg, 2006.

[9] Mendes, P.N., Jakob, M., Garcia-Silva, A., Bizer, C.: DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems, ACM, New York, USA, 1-8, 2011.

[10] Polyvyanyy, A. (2007). Evaluation of a novel information retrieval model: eTVSM. Master's thesis, Hasso Plattner Institut.

[11] Bird, Steven, Edward Loper and Ewan Klein (2009). NLP with Python. O'Reilly Media Inc.

[12] Adam Berenzweig, Beth Logan, Daniel P. W. Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. Computer Music Journal, 28(2):6376, Summer 2004.