Yasavur U, Amini R, Lisetti C, Rishe N: Ontology-based Named Entity Recognizer for Behavioral Health. In: Proceedings of FLAIRS 2013. Association for the Advancement of Artificial Intelligence (www.aaai.org), 2013.

Ontology-based Named Entity Recognizer for Behavioral Health

Ugan Yasavur, Reza Amini, Christine Lisetti, Napthali Rishe
School of Computing & Information Sciences
Florida International University
Miami, FL, 33199, USA

Abstract

Named-Entity Recognizers (NERs) are an important part of information extraction systems in annotation tasks. Although substantial progress has been made in recognizing domain-independent named entities (e.g. location, organization and person), there is a need to recognize named entities for domain-specific applications in order to extract relevant concepts. Due to the growing need for smart health applications in order to address some of the latest worldwide epidemics of behavioral issues (e.g. over eating, lack of exercise, alcohol and drug consumption), we focused on the domain of behavior change, especially *lifestyle change*. To the best of our knowledge, there is no named-entity recognizer designed for the lifestyle change domain to enable applications to recognize relevant concepts. We describe the design of an ontology for behavioral health based on which we developed a NER augmented with lexical resources. Our NER automatically tags words and phrases in sentences with relevant (lifestyle) domain-specific tags (e.g. [un/]healthy food, potentially-risky/healthy activity, drug, tobacco and alcoholic beverage). We discuss the evaluation that we conducted with with manually collected test data. In addition, we discuss how our ontology enables systems to make further information acquisition for the recognized named entities by using semantic reasoners.

Introduction

It has been recently reported that three lifestyle behaviors – poor diet and lack of exercise, smoking, and alcohol consumption – are leading causes of death (main cause for 38% of deaths in US) (Mokdad et al. 2004). However, on the positive side, it is known that that "among U.S. adults, more than 90% of type 2 diabetes, 80% of CAD, 70% of stroke, and 70% of colon cancer are potentially preventable by a combination of nonsmoking, avoidance of overweight, moderate physical activity, healthy diet, and moderate alcohol consumption" (Willett 2002).

In order to address recent epidemic behavior related health problems, traditional hospital-centric medicine is transforming to preventive medicine which focuses on well-being and quality of life. Progress in information technologies will enable the development of smart-health applications which are envisioned to support health-care transformation.

Smart health applications that use text are numerous and the ability to identify behavioral-related concepts in text will therefore support smart-health applications in a variety of ways, including information extraction, retrieval, reasoning tasks, and dialog-based health systems.

Because identifying behavioral concepts in text requires some world knowledge, we created a behavioral health ontology to model world knowledge for behavioral health problems, and a named-entity recognition system based on the ontology.

Traditionally, a named-entity recognition (NER) task focuses on finding and tagging proper nouns into predefined set of classes such as location, organization or person (Tjong Kim Sang and De Meulder 2003). In addition to these mentioned named-entities, recognizing numerical and temporal entities such as date, time, percentage, money have also been studied by researchers (Finkel, Grenager, and Manning 2005).

By contrast, in order to address our focus on behavioral health, we are interested in extracting information about behavior-related concepts which are generally classified as common nouns. Specifically, our main goal is to classify named-entities into categories that are essential for the design and development of behavior health (Matarazzo 1980) systems, which are mostly focused on lifestyle changes.

We currently classify named-entities with the following category labels: *unhealthy/healthy food, potentially risky/healthy place* and *potentially risky/healthy activity*. In addition to these classifications, we can recognize alcoholic beverages, tobacco products and drugs (narcotics). Our label selection was decided based on the most prevalent behavioral health problems.

Our system uses (1) healthy and unhealthy food labels for behaviors related with *diet*; (2) healthy and potentially-risky activity labels for *exercise* and *alcohol consumption* related behaviors (*activity* may involve alcohol such as partying); (3) healthy and potentially-risky place labels for exercise and alcohol consumption (*place* may have or involve alcohol such as night club); (4) alcoholic beverage label to recognize alcoholic beverages; (5) drug label to recognize *drugs*; (6) and tobacco label to label *tobacco products*.

As discussed on the article, if our system can not find polarized label (e.g., healthy food, healthy activity, and potentially-risky place), it uses neutral labels (e.g., food, activity and place). Thus, our NER tags main behavior-related concepts with the mentioned labels.

In the next section, we discuss latest research conducted in the named-entity recognition field and we compare domain-independent NERs against domain-dependent NERs. We then describe our general approach for the design and development of our behavioral health ontology. Finally we discuss the evaluation of our system and the current results that we obtained on a manually collected data set.

Related Work

Ontology-based named entity recognition, annotation, and information extraction is used successfully in different domains including extracting relevant concepts in biological literature (Muller, Kenny, and Sternberg 2004) and the business intelligence domain (Saggion et al. 2007). In the food domain, Weigand et al. (2012) designed a lexical resource for German, to perform relation extraction for recommending products and assisting online customers. A typical relation type is pairs of food items that are suitable to be consumed together. In addition to the relation types, each food in a relation is classified into healthy and unhealthy categories. This system takes also into account context-dependent healthiness (i.e. having a medical condition such as allergy) which requires background information about a user.

Although ontology-based approach can be seen similar to using gazetteers (list of names of entities) approach in NER, the ontology approach provides additional advantages in terms of making further reasoning and knowledge acquisition for extracted concepts. We will discuss it in more detail in the following sections. In addition to using ontology and modeling knowledge using the Web Ontology Language (OWL) (Mcguinness and van Harmelen 2004), we have augmented it with WordNet (Miller 1995). WordNet is used, if a concept does not exist in our ontology.

Traditional domain-independent named-entity recognition mainly concentrates on using supervised techniques to classify proper nouns into small number of predefined categories (Tjong Kim Sang and De Meulder 2003), (Nadeau and Sekine 2007). The disadvantage of this method is collecting and gathering hundreds of labeled training data. Although there is available data for common categories (e.g. location, organization and person), for domain-specific categories it is not the case. Collecting and labeling hundreds of training data is not feasible for domains which deal especially with common names. Because common names (e.g., apple, gym, whiskey) do not have specific word-level features (i.e., orthographic information, orthographic patterns) as proper names (e.g., Apple, IBM, Henry Ford, 3M) which are used widely in supervised systems (Nadeau and Sekine 2007), the feature space for common names is a lot more restricted than proper names' feature space.

Also document and corpus features including multiple occurrences, local syntax, and corpus frequency are not really useful for common names. Although these aspects are disadvantages in terms of using supervised techniques in recognizing common names, there is a possibility to use alternative approaches (e.g. lexical semantic networks, lexical ontologies) which are not directly available for proper name recognition. Because there is no dictionary or lexical resource containing all proper names which are constantly being created.

Moreover, Krupka and Hausman (1998) showed that using extensive gazetteers for proper name recognition does not really improve the recognition accuracy. The advantage of our ontology-based augmented approach is twofold. First, it is not required to build and maintain extensive gazetteers because the system uses semantic network structure based on WordNet. WordNet-based named entity recognition has been implemented successfully for domain-independent NER by extracting trigger words from WordNet (Magnini et al. 2002), and in video annotation applications based on semantic similarity (Qiu, Guan, and Feng 2010). Second, our system's domain is easily modifiable which makes it ontology-dependent but domain-independent. Therefore, we have adopted a different approach based on extendable ontology model which is augmented with the WordNet. Although our NER can be only used for applications that are focused on lifestyle change, it is possible to use the application in different domains by changing the domain of underlying ontology.

Our ultimate goal is to develop an autonomous dialog system for the lifestyle change domain and we are planning to use our NER for common names. Recognizing and classifying domain specific entities from utterances is the first step towards our the goal. Having OWL ontology introduces additional possibilities for the recognized entities by using reasoners to classify them into further categories which has crucial importance in autonomous agent-based dialog systems.

Approach

Ontology Design

We have designed behavioral health ontology in Protégé (Knublauch et al. 2004) with OWL 2. A small snapshot of our ontology is presented in Figure 1.There are 6 top level classes which correspond to categories of named entities. In total there are 60 classes, 490 individuals, and 20 object properties in our current behavioral health ontology.

Healthiness, unhealthiness and potentially-riskiness are defined to address most prevalent risky health behaviors such as obesity, excessive alcohol consumption, drug and tobacco use. For categorizing foods into the healthy and unhealthy categories, we considered following aspects:

- Sugar and calorie content of the food item.
- Fat content of the food item.
- Artificial additives and extensive industrial processing of the food item.

For example broccoli is an instance of the vegetable class which is the subclass of the healthy food class and chocolate is an instance of the snack class which is subclass of the unhealthy food.

Activity and *place* concepts are defined in terms of healthy/unhealthy food, alcoholic beverage, tobacco and drug concepts. We defined relationships between the ontology concepts by using the object properties such as *Have-*

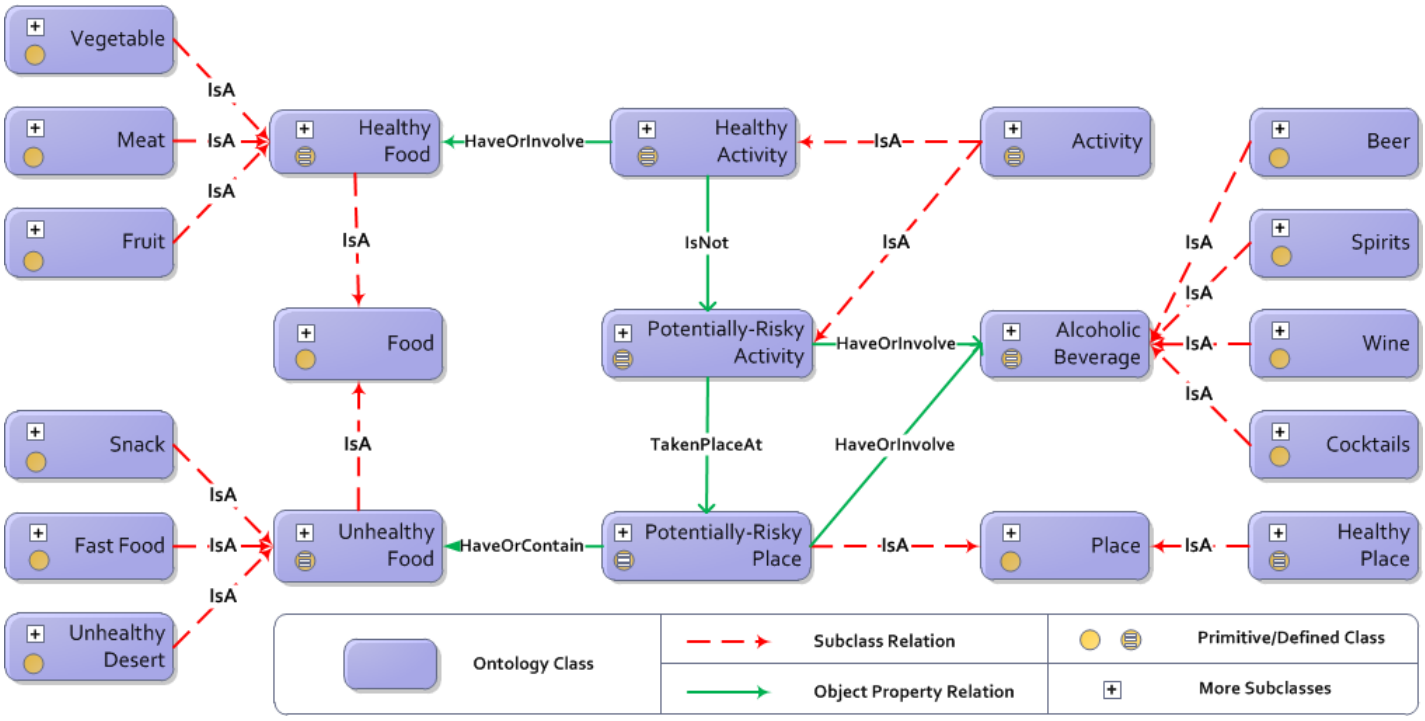

Figure 1: Behavioral Health Ontology

OrInvolve and *TakenPlaceAt* (see Figure 1). As an example, a *potentially-risky activity* (e.g., drinking alcoholic beverage, eating junk food, smoking) may have or involve *unhealthy food* (e.g., fast food), *alcoholic beverage* (e.g., vodka) or *tobacco products*. We also defined the object properties which allows to perform knowledge acquisition between the healthy/potentially-risky activity and place ontology classes. As an example potentially-risky activity assumed to be taken placed at a potentially-risky place (fast food restaurant). For a subset of the ontology structure which shows the relationships between ontology concepts (see Figure 1).

Individuals in the ontology structure represent instances of each class. For example, **grape** is an individual of **fruit** class and transitivity between class structures implies that **grape** is also individual of **healthy food** class (fruit is a subclass of healthy food). In addition to entities with common names, for some classes we include proper names which are frequently used for some ontology classes. For example, **Burger King** is an individual of **fast food restaurant** class which is a subclass of **potentially-risky place** class because fast foods are generally classified as unhealthy. The alcoholic beverage class also contains many instances which has proper names (e.g. beer, vodka, whiskey brands).

We have also defined anonymous classes based on relationships between concepts using object properties. Object properties such as "haveOrInvolve" allow our system to make some inferences, including inferences which are not directly indicated based on the class hierarchy. For example "having or involving alcoholic beverage" is explicitly specified as potentially-risky activity, if we query our ontology by using OWL description logic (DL) query with "drinking *some* Jack Daniels", it can infer that Jack Daniels is a whiskey, whiskey is a spirit, spirit is an alcoholic beverage, and using alcoholic beverage is a potentially-risky activity. Although it is not required for NER task, for applications that require additional information about the recognized entities, our ontology structure can be queried to retrieve taxonomic information about the entities.

WordNet

WordNet is a lexical database of English (Miller 1995). All word groups including nouns are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interconnected by conceptual-semantic and lexical relations. WordNet can be used as a lexical ontology based on hypernym/hyponym relationships between noun synsets. These relationship structure can be interpreted as super-class and subclass relationship as in ontology classes.

WordNet Distance: Semantic distance, similarity, and semantic relatedness are being used interchangeably by researchers and used in annotation, word sense disambiguation, information extraction, information retrieval, etc. Since, there are different measures proposed for relatedness or distance (Pedersen, Patwardhan, and Michelizzi 2004), it is important to distinguish these terms.

Budanitsky and Hirst (2006) distinguish **semantic relatedness** as a more general concept of **similarity**. They attempt to demonstrate the difference between relatedness and similarity by an example: "Similar entities are semantically related by virtue of their similarity (bank-trust-company), but

dissimilar entities may also be semantically related by lexical relationships such as meronymy (car-wheel) and antonymy (hot-cold), or just by any kind of functional relationship or frequent association (pencil-paper, penguin-Antarctica, rain-flood)." Therefore, similarity and relatedness does not refer to the same concept.

The semantic distance term generates even more confusion in terms of relatedness and similarity. Therefore, there are different approaches to calculate it. The semantic distance we are referring to is the distance in hypernym/hyponym tree. As we have mentioned before, WordNet can be interpreted as an ontology based on hypernym/hyponym relations. Thus distance between two words in hypernym/hyponym tree is more compatible with our goals for NER than relatedness or similarity concepts.

We used RiTa.WordNet[1] library to calculate semantic distance. The algorithm calculates the distance between any two senses of the two words (results is normalized within 0-1) with the specified Part-Of-Speech(POS) tag. For our purposes we use noun as POS tag. The algorithm (1) finds common parents of the two words, (2) calculates the minimum distance (shortest path) to the common parent from either of the words, (3) calculates the distance from the common parent to root of tree, and (4) normalizes the result (see Algorithm 2).

Architecture

The NER performs tasks to locate nouns in the sentences based on the output of the Stanford Part-Of-Speech Tagger, then the identified nouns are lemmatized with the lemmatizer available in Stanford CoreNLP tool (Toutanova et al. 2003). The identified and lemmatized nouns are passed to the Tagger algorithm to be labeled into the following categories: (1) Healthy Food; (2) Unhealthy Food; (3) Healthy Activity; (4) Potentially-risky Activity; (5) Healthy Place; (6) Potentially-risky Place; (7) Drug; (8) Alcoholic Beverage; and (9) Tobacco. The system may use neutral labels from ontology for tagging, if the system can not identify polarized label for the named-entity (e.g. instead of unhealthy food, food label can be used).

Tagger algorithm (1) queries classes in the ontology, if it finds a matching class, it traverses the ontology to higher level classes to find a appropriate tag; (2) if the lemma is not equal to the name of any classes, it queries individuals in the ontology and finds the class of an individual (if the individual exists) and traverses the ontology to find an appropriate label; (3) a) if the noun does not exist in the ontology, it uses the distance algorithm (see Algorithm 2), the ranker component (see Figure 2) compares the distance between each class and the parameter noun, and then the tagger algorithm selects the class with the minimum distance to the noun; b) if the selected class is a first level class (e.g. Alcoholic Beverage, Drug/Narcotic) and the distance is less than the threshold distance, it tags the name with the corresponding label; c) if the selected class is in lower position (e.g. Beer, Cannabis) in the hierarchy and the distance is less than the threshold distance, it tags with the corresponding tag.

[1] http://www.rednoise.org/rita/wordnet

The intuition behind using different distances for different level classes is as follows: if the minimum distance of a parameter noun is calculated for a lower level class in the ontology, it is expected that parameter noun is also a specific term, so the minimum distance to their common parent in hypernym/hyponym tree is expected to be short. For example, if the noun is *margarita* and the closest ontology class is *martini*, the expected distance is short because *martini* is a low level class in ontology. If *martini* did not existed in ontology and shortest distance to *margarita* is from *alcohol* class, the expected distance is longer than the distance of specific class because *alcohol* class is a high level class. For this specific example the distance between *margarita* and *martini* is 0.1 and their common parent is *cocktail* in tree hierarchy. The distance between *margarita* and *alcohol* is 0.3 where alcohol is the parent of margarita (common parent is alcohol too). Therefore using different threshold values for different level classes in the ontology helps to fine tune coverage of extension based on WordNet.

Algorithm 1 Tagger

if Is parameter noun(**pn**) a class in ontology **then**
  Tag the noun with super class of corresponding class
else if Is **pn** individual in ontology **then**
  1.Find class Of the individual
  2.Find super class of individual's class
  Tag the **pn** with the super class
else
  1.Compare minimum distance between noun and Ontology classes by using **Distance** algorithm (See Algorithm 2)
  2.Select the class with shortest path to the noun
  3.If Selected class is first level and distance is less than higher-threshold
  return it as Tag
  4.If selected class is not first level and distance is less than lower-threshold
  return it as Tag
  Otherwise do not tag
end if

Data and Evaluation

Since there was no tagged data in our domain, we collected the test data manually from a variety of related websites which have relevant domains. For example, we have used meal recipe websites to find data related with the food domain. We have collected 88 sentences with 220 named-entities. Two annotators tagged the collected test data with the aforementioned labels. Then we performed two experiments with our NER. The recognition of an entity without healthy, unhealthy or potentially-risky label for the food, activity and place entities considered wrong. For example, if an entity recognized as *food* without having healthy/unhealthy label, we did not count it as a correct recognition. Since the *alcoholic beverage*, *drug* and *tobacco* ontology classes are all considered as unhealthy or potentially-risky in terms of behavioral health, the recognition of an entity in these cate-

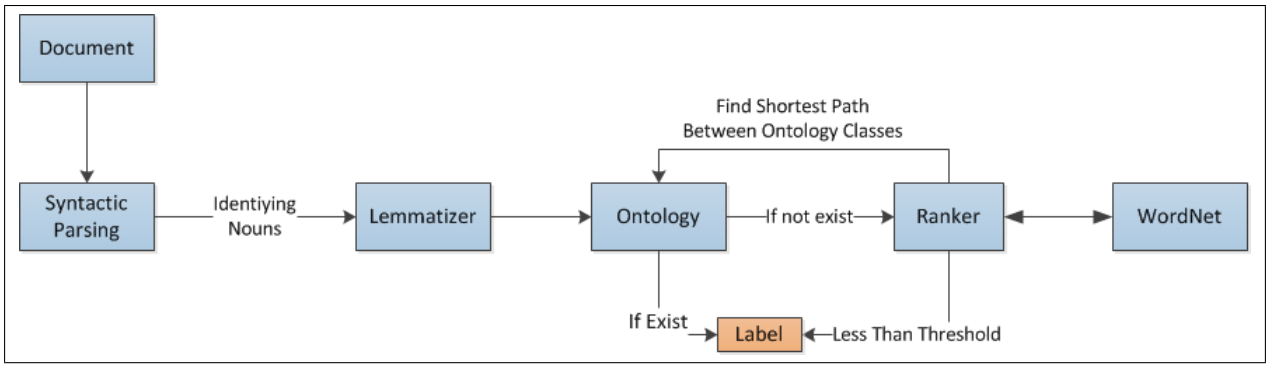

Figure 2: The System Architecture

Distance	Precision	Recall	$F_{\beta=1}$
0.1 and 0.2	83.32%	71.28%	76.80%
0.2 and 0.3	65.55%	81.44%	72.64%

Table 1: Overall precision, recall and $F_{\beta=1}$ rates obtained by conducting two experiments with different threshold distances.

gories does not need an additional label.

The experiment results are presented in Table 1. The first experiment was conducted using 0.2 as the distance threshold to the high level classes and 0.1 distance threshold the low level classes in the ontology. Second experiment was conducted using 0.3 as the distance threshold to the high level classes and 0.2 distance threshold to the low level classes. The performance of the NER is measured with $F_{\beta=1}$ rate:

$$F_\beta = \frac{(\beta^2+1) * precision * recall}{\beta^2 * precision + recall}$$

with β=1 (Rijsbergen 1979). Where, *Precision* is the percentage of the named entities which are correctly recognized by the system and *Recall* is the percentage of the named entities present in the test data that are recognized by the system. A recognized named entity is correct only if it is the exact match of the corresponding entity in the manually tagged file. First, we conducted the experiment with the higher distance threshold values, so, the precision was low because of the false positives (unexpected results). The high number of false positive results were caused by the high threshold distances in the WordNet tree. We encountered many problems due to the word-sense ambiguity. For example *ice* and *glass* words were labeled as drug because *ice* and *glass* as a slang refer to a kind of drug. We observed many similar problems to this example in the first experiment.

In the second experiment, the precision increased significantly while recall decreased. It was the result of the low threshold values for the distances. In this experiment the number of false positives decreased significantly while false negatives (missing results) increased. It was the result of the decreased coverage of the system due to the low distance threshold. We did not observe as many unexpected results as in the first experiment because of the slangs but we observed an increase in unlabeled named-entities.

Another factor which affects the results is the output of the part-of-speech tagger and lemmatizer. Although they worked with high accuracy in general, for some cases they did not give expected output.

Overall, although precision, recall and F-Measure results are not high, the results are acceptable for the first version and promising for the future versions.

Algorithm 2 Distance Algorithm

Locate the common parent of the two lemmas by checking each sense of each lemma
if No common parent found **then**
  return 1
else
  1.Calculate min distance to common parent (the shortest path from either lemma to common parent)
  2.Calculate distance from common parent to root (length of the path from common parent to the root of WordNet ontology)
  3.Calculate and return the normalized distance to common parent as:
  (*minDistToCommonParent* / (*distFromCommonParentToRoot* + *minDistToCommonParent*))
end if

Conclusion

We designed a named-entity(NE) recognizer for the lifestyle change domain. We addressed the differences between traditional NE recognition and the domain specific NE recognition. To address our problem in recognizing lifestyle related entities in text, we designed a behavioral health ontology. Based on our ontology model, we created a named entity recognizer. Also, we identified other possible use-cases of our ontology. To extend the ontology for the named-entity recognition purposes, we augmented it with the WordNet. We used a hypernym/hyponym tree and calculated distances between synsets.

We conducted two experiments with different distance threshold values and reported the results. We observed that

threshold distance has a significant effect on precision and recall. While high threshold values increase the recognition rate, it causes unexpected false positives because of wrong labels. We believe that we can address this issue by using dynamic distance threshold for different ontology classes in future. Although resulting precision, recall and F-Measure results are not high, they are acceptable for the first version and promising for the future versions. In the future versions of our named-entity recognizer, we will conduct experiments with the dynamic distance thresholds.

References

Budanitsky, A., and Hirst, G. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.* 32(1):13–47.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 363–370. Stroudsburg, PA, USA: Association for Computational Linguistics.

Knublauch, H.; Fergerson, R. W.; Noy, N. F.; and Musen, M. A. 2004. The Protégé owl plugin: An open development environment for semantic web applications. 229–243. Springer.

Krupka, G., and Hausman, K. 1998. IsoQuest Inc.: Description of the NetOwl (TM) Extractor System as Used for MUC-7. *In Proceedings of the Seventh Message Understanding Conference (MUC-7)* 1–10.

Magnini, B.; Negri, M.; Prevete, R.; and Tanev, H. 2002. A wordnet-based approach to named entities recognition. In *Proceedings of the 2002 workshop on Building and using semantic networks - Volume 11*, SEMANET '02, 1–7. Stroudsburg, PA, USA: Association for Computational Linguistics.

Matarazzo, J. D. 1980. Behavioral health and behavioral medicine: Frontiers for a new health psychology. *American psychologist* 35(9):807.

Mcguinness, D. L., and van Harmelen, F. 2004. OWL web ontology language overview. W3C recommendation, W3C.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38:39–41.

Mokdad, A. H.; Marks, J. S.; Stroup, D. F.; and Gerberding, J. L. 2004. Actual causes of death in the United States, 2000. *JAMA : the journal of the American Medical Association* 291(10):1238–45.

Muller, H.-M.; Kenny, E. E.; and Sternberg, P. W. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2(11):e309.

Nadeau, D., and Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3–26. Publisher: John Benjamins Publishing Co.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, 38–41. Stroudsburg, PA, USA: Association for Computational Linguistics.

Qiu, Y.; Guan, G.; and Feng, D. 2010. Improving News Video Annotation with Semantic Context. *2010 International Conference on Digital Image Computing: Techniques and Applications* 214–219.

Rijsbergen, C. J. V. 1979. *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 2nd edition.

Saggion, H.; Funk, A.; Maynard, D.; and Bontcheva, K. 2007. Ontology-based information extraction for business intelligence. In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, 843–856. Berlin, Heidelberg: Springer-Verlag.

Tjong Kim Sang, E. F., and De Meulder, F. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, 142–147. Stroudsburg, PA, USA: Association for Computational Linguistics.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *IN PROCEEDINGS OF HLT-NAACL*, 252–259.

Wiegand, M.; Roth, B.; Lasarcyk, E.; Köser, S.; and Klakow, D. 2012. A gold standard for relation extraction in the food domain. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, volume 51.

Willett, W. C. 2002. Balancing life-style and genomics research for disease prevention. *Science (New York, N.Y.)* 296(5568):695–8.