4. Experiments

This chapter presents our experimental results using the features described in Chapter 3. Section 4.1 describes the corpus and the ontology used in this dissertation. Section 4.2 gives the experimental setup defining all the algorithms used for training and for feature extraction. Section 4.3 presents our results for the held-out experiments. Finally, Section 4.4 presents our results for the human evaluation experiments.

4.1. Corpus

In this dissertation, we adopt DBpedia to populate our triple store. DBpedia is a project that derives a data corpus from Wikipedia. The Wikipedia project is a free and collaboratively edited encyclopedia which is available in over 250 languages, with the English one accounting for more than 1.95 million articles. All these data is accessible on the Web under the terms of the Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License.

Apart from the task of parsing structured data from Wikipedia, there is another effort for creating an ontology that conceptualize the data in the DBpedia dataset. This ontology is called the DBpedia Ontology. It organizes all the structured data from Wikipedia into classes and provides semantics for the data extracted (Sahnwaldt, 2012) classifying concepts into 359 classes organized into hierarchies. The DBpedia ontology currently contains over 2,350,000 instances and more than 480 different relations. DBpedia is the largest dataset present in the LOD Cloud.

As a source of unstructured text, we used all Wikipedia articles in English. We annotated Wikipedia articles with entities from DBpedia by matching links to

others articles in the text to entities in the DBpedia. There is a correspondence between every entity in DBpedia to each unique article in Wikipedia. By doing this, there is no ambiguity in the annotation of DBpedia entities since links in Wikipedia articles were created by the editors of the articles.

Hence, we consider a sentence in a Wikipedia article as *applicable* to our propose if it contains at least two links to other articles, which can be easily translated into DBpedia URIs. For sentence boundary detection, we used the algorithm proposed by Gillick (2009).

We applied two heuristics to increase the number of applicable sentences. The first heuristic attempts to annotate references to the main subject of an article. For example, the Wikipedia article about *Barack Obama* has no self links, therefore there will be no annotations about the *Barack Obama* entity available in DBpedia in the article. Assuming that the article contains valuable information about *Barack Obama*, triples about the main subject would not be extracted if no treatment is done. For that reason, we annotate every match between the article text and the article title.

We also identify proper names by simply looking for capital letters at the beginning of tokens in the article title and we give them a different treatment. The article about *Barack Obama*, for example, contains several mentions of the main subject as *Obama* instead of his complete name. So, for proper names, we match every token that composes the proper name to the article subject.

The second heuristic explores sentences with more than two instances annotated. We intend to explore combinations of such annotations. All the combinations of pairs of instances are taken as examples to be used. For example, the sentence "***Obama*** *is a graduate of* ***Columbia University*** *and* ***Harvard Law School"*** were explored to generate three applicable sentences:

1. ***Obama*** *is a graduate of* ***Columbia University*** *and Harvard Law School*

2. ***Obama*** *is a graduate of Columbia University and* ***Harvard Law School***

3. Obama is a graduate of ***Columbia University*** *and* ***Harvard Law School***

Applying all the strategies described above, we generated a corpus of over 2.2 million sentences with annotated entities.

4.2. Experimental Setup

From the corpus described in Section 4.1, for lexical features extraction, we used the Stanford Part of Speech Tagger (Toutanova et al., 2000) and the WSJ 0.18 Bidirectional model for POS features. We also simplified the POS tags into nine categories: nouns, verbs, adverbs, prepositions, adjectives, numbers, foreign words, possessive ending and everything else. We extracted a total of 2,276,647 sentences. From them, we extracted feature vectors that were used as input for the Logistic Regression classifier.

Figure 18 shows the distribution of examples for each relation. The x-axis represents each of the 480 relations. They are in decreasing order by the number of occurrences represented by the y-axis. The first relation, for example, has 607,308 examples and this number decays to 159,717 for the second relation. The abrupt decay of the number of examples by relation can be noticed in Figure 18 in the interval $0 \leq x \leq 100$. The top 20 relations in the number of examples are shown in Table 3.

Our experiments are divided into two classes. The first one is the held-out experiments which comprehend tests where part of the data is held out for testing and the remaining is used for training a classifier. This test is intended to find out how well the classifier performs, but this is not a perfect test since our dataset was heuristically labeled. For example, consider the following two sentences with two annotated entities *Robbie* and *Suzie*:

Robbie is **Suzie**'s husband**.**

Robbie worked with **Suzie** on the X-project.

Suppose that *Robbie* and *Suzie* are instances of one ontology and are only related by the predicate *married_with.* Heuristically, the two sentences above would be inserted in our dataset as examples of the relation *married_with.* The first sentence is a valid example but the second sentence is not.

Also, sentences may not be classified correctly in agreement with thematic roles of annotated entities in the sentence. For example, consider the relation *murder* and the following two sentences:

Robbie killed **Suzie** and his friend**.**

Robbie and **Suzie** killed his friend**.**

In the first sentence, *Robbie* is the agent of the phrase and *Suzie* is a victim but on the second sentence *Suzie* is also an agent. Although both sentences are examples of the relation *murder,* the thematic roles of each instance would invalidate the second sentence as an example of the relation *murder* considering the entities *Robbie* and *Suzie* since they are co-agents. We did not apply any thematic role processing on our dataset labeling step. Winston (1977) extensive discusses how to identify thematic roles in sentences.

Therefore, to address heuristic issues, the second class of experiments is the human evaluations with the most popular relations of our dataset. By doing so, we calculate the accuracy for the most popular relations in our dataset and therefore obtain a more accurate evaluation of the approach proposed in this dissertation.

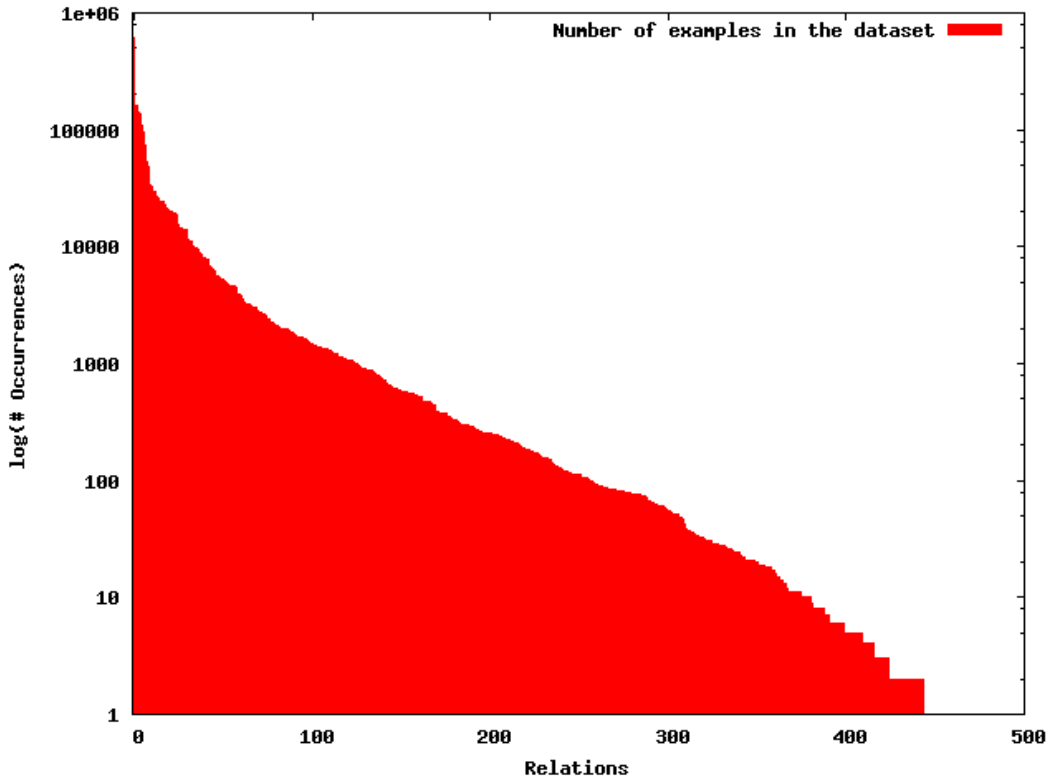

Figure 18: Histogram for the number of occurrences for each relation

Table 3: Top 20 relations in number of examples in the dataset

Relation	Number of examples
http:/dbpedia.org/ontology/country	607,380
http:/dbpedia.org/ontology/family	159,717
http:/dbpedia.org/ontology/isPartOf	139,694
http:/dbpedia.org/ontology/birthPlace	138,797
http:/dbpedia.org/ontology/genre	109,813
http:/dbpedia.org/ontology/location	96,516
http:/dbpedia.org/ontology/type	72,942
http:/dbpedia.org/ontology/order	53,421
http:/dbpedia.org/ontology/occupation	48,859
http:/dbpedia.org/ontology/hometown	34,010
http:/dbpedia.org/ontology/state	33,199
http:/dbpedia.org/ontology/region	29,577
http:/dbpedia.org/ontology/genus	29,177
http:/dbpedia.org/ontology/artist	26,784
http:/dbpedia.org/ontology/deathPlace	26,184
http:/dbpedia.org/ontology/class	24,698
http:/dbpedia.org/ontology/language	24,186
http:/dbpedia.org/ontology/associatedMusicalArtist	22,627
http:/dbpedia.org/ontology/city	21,330
http:/dbpedia.org/ontology/spokenIn	20,965

4.3. Held-Out Evaluation

For the held-out evaluation experiments, half of the sentences for each relation were randomly chosen not to be used in the training step. They are later used in the testing step.

We run experiments using only lexical features, semantic features and both set of features so it is possible to measure the impact of the proposed features in this dissertation. The comparison between results are made by counting the number of classes which the classifier responds with F-measure greater than 70%. The top 20 classes are displayed in tables and class names are displayed only using it's URI prefix. They share the same prefix: *http://dbpedia.org/ontology*.

We considered as baseline the number of classes with F-measure greater than 70% that a classifier trained only with lexical features respond. This set of results are shown in Table 4. Our baseline has a total of nine classes.

Table 4: Relations for a classifier trained with lexical features only

Class	Precision	Recall	F-measure
/targetSpaceStation	1	1	1
/department	0.98	0.86	0.92
/discoverer	1	0.81	0.9
/militaryBranch	0.94	0.83	0.88
/notableWine	0.99	0.75	0.85
/programmeFormat	0.87	0.77	0.82
/type	0.69	0.83	0.75
/license	0.98	0.58	0.73
/sport	0.81	0.63	0.71

The next experiment considered the computation of a classifier using only semantic features proposed in this dissertation. We achieved a total of 60 classes with F-measure greater than 70%. The top 20 classes are shown in Table 5. The total number is more than six times bigger than our the baseline.

The classes */department* and */license* are presented in Table 4 but they were the only ones not classified with more than 70% of F-measure by the classifier trained with semantic features.

The final experiment considered both lexical and semantic features. We obtained a total of 88 classes with F-measure greater than 70%. Thus, a gain of 31.82% was achieved by combining the two set of features. The top 20 classes are shown in Table 6.

Table 5: Top 20 relations for a classifier trained with semantic features only

Class	Precision	Recall	F-measure
/areaOfSearch	1	0.98	0.99
/ground	0.96	1	0.98
/mission	0.97	1	0.98
/politicalPartyInLegislature	1	0.95	0.97
/precursor	0.99	0.96	0.97
/sport	0.96	0.97	0.97
/targetSpaceStation	0.94	1	0.97
/discoverer	0.93	1	0.96
/drainsTo	0.97	0.93	0.95
/isPartOfAnatomicalStructure	0.91	1	0.95
/ideology	0.92	0.95	0.94
/academicDiscipline	0.88	0.99	0.93
/locatedInArea	0.9	0.97	0.93
/notableWine	0.93	0.92	0.93
/programmeFormat	0.88	0.99	0.93
/followingEvent	1	0.85	0.92
/fuel	0.85	1	0.92
/musicalBand	0.85	0.98	0.91
/originalLanguage	1	0.81	0.9

Although there is a considerable gain of performance by using both sets of features, several classes predicted by only semantic features have a worse performance when adding lexical features. For example, the class */aircraftFighter* with F-measure of 77% on the semantic feature based classifier has F-measure of 50% on the classifier with both features. This is an example of the degradation of performance that classes receive when adding lexical features.

However, our experiments showed that the number of classes that are classified with more accuracy and better recall by adding lexical features is greater than the degraded ones. To summarize, Table 7 compares the use of set of features, expressing our gain of almost ten times over the baseline.

Table 6: Top 20 relations for a classifier trained with lexical and semantic features

Class	Precision	Recall	F-measure
/areaOfSearch	1	0.97	0.98
/ground	0.97	1	0.98
/mission	0.99	0.96	0.97
/sport	0.97	0.97	0.97
/targetSpaceStation	1	0.93	0.97
/academicDiscipline	0.93	0.99	0.96
/discoverer	0.99	0.93	0.96
/locatedInArea	0.93	0.98	0.96
/programmeFormat	0.93	0.99	0.96
/politicalPartyInLegislature	1	0.91	0.95
/precursor	0.99	0.91	0.95
/team	0.94	0.95	0.95
/drainsTo	0.9	0.98	0.94
/department	0.97	0.89	0.93
/fuel	0.93	0.93	0.93
/musicalBand	0.89	0.97	0.93
/statisticLabel	0.87	0.99	0.93
/isPartOfAnatomicalStructure	0.88	0.95	0.92
/notableWine	0.97	0.87	0.92

Table 7: Number of classes with at least 70% of F-measure by set of features

Features	Number of classes > 70% F-measure
Lexical	9
Semantic	60
Lexical + Semantic	88

4.4. Human Evaluation

For the human evaluation, the same separation of the examples into training/testing as defined for the held-out experiments set was performed. Half of the sentences for each relation were randomly chosen not to be used in the training step. From the remaining sentences, random samples of 100 sentences were extracted from each of the top 10 relations in the number of examples in our

dataset. Those samples were forwarded to two evaluators. The evaluation of the accuracy of each prediction of the samples was carried out manually. The results are showed in Table 8.

Table 8: Average accuracy for the top 10 relations in examples in our dataset for human evaluation of a sample of 100 predictions.

Relation	Average Accuracy
http:/dbpedia.org/ontology/country	0.73%
http:/dbpedia.org/ontology/family	0.75%
http:/dbpedia.org/ontology/isPartOf	0.9%
http:/dbpedia.org/ontology/birthPlace	0.76%
http:/dbpedia.org/ontology/genre	0.77%
http:/dbpedia.org/ontology/location	0.76%
http:/dbpedia.org/ontology/type	0.8%
http:/dbpedia.org/ontology/order	0.81%
http:/dbpedia.org/ontology/occupation	0.87%
http:/dbpedia.org/ontology/hometown	0.68%

Samples consist of sentences with explicit entity annotations. An example of a sentence is:

<Camel/Camel_(band)> are an <English/England> progressive rock band formed in 1971.

In the above sentence, there are two annotated instances. The first one refers to the instance "http://dbpedia.org/resource/Camel_(band)" and is referenced in the sentence by the word *Camel.* The second entity is "http:// dbpedia.org/resource/England" and is referenced by the word *English.*

The sentence above is an example of an issue of the human evaluation process. Words that represent entities and the entities themselves are important parts of the interpretation of the relation. In the above example, if we consider that Camel is an English progressive rock band, we could assign a relation of, for example, http://dbpedia.org/ontology/type. However, if we consider that the word *English* refers to the entity that represents the country England, a different relation can be assign, such as http://dbepdia.org/ontology/hometown, which is the one pointed by our classifier. The evaluators must be aware of the possibility of such interpretations.

Another issue of human evaluation is that there are relations that required domain specific knowledge such as http://dbpedia.org/ontology/family and http://dbpedia.org/ontology/order. Two respective examples of such relations follow, indicating a very specific vocabulary of the biological taxonomy domain:

<Gymnothorax/Gymnothorax> is a genus of <moray_eels/Moray_eel> in the family Muraenidae.

Vanellus is the genus of <waders/Wader> which provisionally contains all <lapwings/Lapwing> except Red-kneed Dotterel, "Erythrogonys cinctus".

Finally, another issue of the evaluation process is the definitions of the relations. Relations, like http://dbpedia.org/ontology/birthPlace and http://dbpedia.org/ontology/hometown without the context of a sentence, can almost be used interchangeably, making the validation of isolated sentences for those relations ambiguous.

Almost every relation on our results presented an average accuracy greater than 70%, except http://dbpedia.org/ontology/hometown, with 68%. The best accuracy were achieved by the relation http://dbpedia.org/ontology/isPartOf, with 90%.

4.5. Summary

This chapter presented our experimental results, including a brief discussion about the tools and corpus used. Two types of experiments were performed: held-out experiments and human evaluation experiments. The held-out evaluation showed the impact of the semantic feature proposed in the work over the lexical features. We concluded that the combination of our semantic feature and lexical features achieves the best results, obtaining a total of 88 classes with more than 70% of F-measure. The human evaluation experiments indicated the average accuracy for the top 10 relations in number of examples in our dataset. Despite of the issues on evaluating some relations, the majority of the classes achieved an average accuracy greater than 70%.