

INSA LYON
UNIVERSITÄT PASSAU

MASTER THESIS

Image Annotation Network

Author:
Mael OGIER

Supervisors:
Dr. David COQUIL
Dr. Elöd EGYED-ZSIGMOND

*This thesis submitted in fulfilment of the requirements
for the degree of Master of Science*

in the

Informatique - Information und Kommunikation (IFIK)
Double Master Program

August 2015

"I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web—the content, links, and transactions between people and computers. A "Semantic Web", which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines."

Sir Tim Berners-Lee

Acknowledgements

This master thesis was done in the context of the double master degree Informatique - Information und Kommunikation (IFIK), which brings together two Master programs : a degree in computer engineering at the National Institute of Applied Sciences in Lyon (INSA Lyon) and a Master in Informatik (Schwerpunkt : Information und Kommunikationssysteme) at the University of Passau.

First, I would like to thanks the initiators of the double Master IFIK for their valuable efforts and commitment to the german-french collaboration : Prof. Dr. Harald Kosch for the german side and Prof. Dr. Lionel Brunie for the french one.

An unmeasurable amout of thanks goes to my advisors, Dr. David Coquil and Dr. Elöd Egyed-Zsigmond for their precious help and advices during this thesis preparation.

I would also thanks my colleagues and friends for their support as well as my five german roommates who made my time in Passau such enjoyable.

Finally, I thanks the Dropkick Murphys for their energizing music and the worldwide coffee producers for this magical beverage they provide us.

INSA LYON
UNIVERSITÄT PASSAU

Abstract

IFIK
Double Master Program

Master of Science

Image Annotation Network

by Mael OGIER

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too. . .

Contents

Acknowledgements	ii
Abstract	iii
Contents	iv
List of Figures	vii
List of Tables	viii
Abbreviations	ix
1 Introduction	1
1.1 Background	1
1.2 Motivation	2
1.3 Thesis Objectives	3
1.4 Thesis Outline	3
I State of the Art	5
2 Semantic Web Resources	6
2.1 Generalities	6
2.2 DBpedia	8
2.3 Geonames	9
2.4 WordNet	11
2.5 ImageNet	12
3 Similarity Measures	13
3.1 Edge-based	13
3.1.1 Shortest Path	13
3.1.2 Wu-Palmer	14
3.1.3 Evolved Wu-Palmer	15
3.2 Node-based	16
3.2.1 Resnik	16
3.2.2 Lin	16
3.2.3 Jiang and Conrath	17

3.3	Statistical similarity	17
4	Disambiguation	18
4.1	How does it work	18
4.2	Practical example : DBpedia Spotlight	18
5	Existing services	19
5.1	Web service 1	19
5.2	Web service 2	19
5.3	Annotation via stats	20
6	Conclusion	21
6.1	Section 1	21
II	Contribution	22
7	Proposed Methodology	23
7.1	Section 1	23
8	Proposed Architecture	24
8.1	Technology Choices	24
8.1.1	Java	24
8.1.2	Neo4j	24
8.1.3	NLP	25
8.1.4	DBpedia Spotlight	25
8.1.5	JAWS	25
8.1.6	JENA	25
8.1.7	JSoup	26
8.2	Graph Structure	26
8.2.1	Pro-Cons	26
8.2.2	Vertexes	26
8.2.3	Edges	27
9	Experiments	28
9.1	Dataset	28
9.2	Code Explanation	28
9.3	Results and Analysis	29
9.3.1	Evaluation methodology	29
9.3.2	Graph-based experiments	29
9.3.2.1	Direct Neighbors	29
9.3.2.2	Lists - WL	29
9.3.2.3	Lists - SL	29
9.3.3	Plain-text experiments	29
9.3.3.1	WikiLinks	29
9.3.3.2	WikiContent	29
10	Conclusion	30
10.1	Section 1	30

A Appendix Title Here	31
------------------------------	-----------

Bibliography	32
---------------------	-----------

List of Figures

3.1 Sample ontology	13
-------------------------------	----

List of Tables

Abbreviations

LCS	Lowest Common Subsumer
RDF	Resource Description Framework
SP	Shortest Path
SPARQL	SPARQL Protocol and RDF Query Language
URI	Universal Resource Identifier
WP	Wu-Palmer measure
WP++	Wu-Palmer evolved measure
XML	eXtensible Markup Language

Chapter 1

Introduction

1.1 Background

Image is a popular medium nowadays : it is easy to capture, can be really light on your computer and speaks to everyone without distinction of language.

In the all days life, people share their pictures on social networks in less than a blink of eye. In average, 70M of pictures are posted on Instagram each day and the users hit the “Like” button 2.5B times¹. Other services like Picasa or Flickr exists but aren’t as used as Instagram which is the favorite in the eyes of the teen public.

Companies also produce a lot of media data. Industry companies need their products’ pictures, marketing and advertising studios use a lot of images in order to create new stuff for their client, . . . But the most consumer of media data are obviously mass media themselves : Newspapers, TV shows, news broadcasts are dealing with pictures at every moment of their day.

This huge production and consumption of images implies the need of an efficient way to store and search for the relevant one when the time comes. The best illustration to this need is to think of the nice but long moments one had with its relatives searching for the good picture of the new-born nephew in the family pictures album.

Since an image itself doesn’t have a natural plain-text representation the best way to

¹Stats from : <https://instagram.com/press/>

describe it is to add meta-data (data about the data) such as its date of creation, its dimensions or, and this is what this thesis is about, some tags.

There are a lot of ways if one wants to annotate pictures. We can do it manually, using our own words (like “Dad”, “Home” . . .), we can also analyze the raw picture, its pixel representation and compare some metrics (like the color histogram) to sample images in order to detect known concepts. Moreover, if the image already possesses annotations, we can enrich it semantically.

This field is so wide that it is impossible to speak about all the possibilities and technologies. In this study, we will focus on the last point and investigate the automation of the semantic enrichment. We will study the resources at our disposal and propose a solution keeping in mind the facts cited previously.

In the following section, we will present and discuss an application scenario to illustrate the motivation behind this thesis.

1.2 Motivation

NewsTV is a famous TV news channel which runs 24/7 and only speaks about the current news. It has lot of reporters worldwide, covering the important local news and sending their production to the main site in Paris, France.

The employees often need to consult older coverages in order to explain the context of the news, to make the necrology of a famous actor who recently died or to re-use common shots. Therefor, they need to query the central multimedia database management system using keywords they are familiar with like “Elections, France, 2007, José Bové”. But sometimes, their research aren’t so specific and they are looking for more generic pictures, let say “Land, Tree, Animal”.

The first kind of keywords had been tagged by the former reporter who produced the coverage but he logically didn't think to add generic terms. NewsTV needs something to do it automatically when a picture, or any media, is first added to its system which a couple of initial tags.

Details about which kind of technology can be used to achieve this automatic tagging will come in the following sections. To summarize, the goal of this thesis is to propose a running prototype and evaluate different methods of tagging. The questions that we will try to answer during this study are described in the following section.

1.3 Thesis Objectives

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pel-lentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Cur-abitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

1.4 Thesis Outline

The remainder of this thesis will be organized as follows :

Chapter 2 - Semantic Web Resources: presents different semantic web resources, their structures and how to browse them and how are they used in the literature.

Chapter 4 - Disambiguation: reviews the literature and assesses the most relevant ways to disambiguate a list of keywords which may be organize into sentences or not.

Chapter 3 - Measures: provides a solid background on semantic similarity and distance measures. We explore different metrics illustrating their pro/cons with examples.

Chapter 5 - Existing Services: describes existing image annotation services as well as their use-cases.

Chapter 6 - State of the Art Conclusion: summarizes the findings of the previous state of the art and opens the way to the presented contribution.

Chapter 7 - Proposed Methodology : presents the chosen methodology as well as some organizational points.

Chapter 8 - Proposed Architecture : details the technological choices by comparing them to their competitors and the chosen DBMS schema. Illustration figures will be presented.

Chapter 9 - Proposed Architecture : presents the chosen dataset, details some of the main algorithms and reviews the tests' results with the use of different evaluation methods.

Chapter 10 - Contribution Conclusion : summarize the findings of the presented research problem.

Part I

State of the Art

Chapter 2

Semantic Web Resources

Our study is focus on semantic enrichment of an initial set of keywords which can be organized as sentences or not. It is important to first understand what is a semantic concept and how concepts are organized into ontologies.

In this section, we will present some general notions about semantic concepts and review several semantic resources, their hierarchical structures and how we access them.

2.1 Generalities

Linguistic semantics is the study of meaning that is used for understanding human expression through language. It is easy for two human-being to communicate (given that they speak the same language) and to understand what their partner say even if he's using a tricky turn of phrase. However, this task becomes way more difficult when it comes to the comprehension of the human language by a machine. How can the computer guess that "I am totally dead" means in fact "I am really tired" and that the speaker isn't actually dead ? Machines need structured resources to understand us and the Semantic Web is one of them.

The notion of "Semantic Web" has been mentioned for the first time by Berners-Lee et al in [1]. In this paper, they describe it as a Web which is readable by machines in opposite

of most of Web's content which were designed for humans to read. The Semantic Web isn't a separate Web but an extension of the current one which will bring structure to the meaningful content of Web pages.

Two main technologies are used for the development of the Semantic Web : eXtensible Markup Language (short XML) and the Resource Description Framework (short RDF). XML allows everyone to create their own tags and to arbitrary structure their documents but gives no information about what this structure means. Meaning is provided by RDF which stores it in sets of triples which are composed of a subject, a predicate and an object. Those three components can be related to the subject, the verb and object of an elementary sentence. In [2], Miller present a short introduction to the RDF standard and precise that a "Resource" can be any object which is uniquely identifiable by a Uniform Resource Identifier (URI).

The third basic component of the Semantic Web are collections of information called ontologies. An ontology is, in computer science, a document which defines the relations among concepts. Basically, Web ontologies are composed of a taxonomy, which defines classes of objects and their relations, and a set of inference rules.

In addition, the The New Oxford Dictionary of English defines the notion of "semantic concept" as : *An idea or thought that corresponds to some distinct entity or class of entities, or to its essential features, or determines the application of a term, and thus plays a part in the use of reason or language.*

Given those basic notions, we will now further detail four semantic web resources, their taxonomies and review some of their usage found in the literature.

2.2 DBpedia

DBpedia¹ is a project originally launched by two German universities (Berlin and Leipzig) and backed by an important community. It explores Wikipedia² and extract information from it which results on the creation of a multilingual, large-scale knowledge base. The extraction framework, all the available end-points as well as some facts and figures about the project are presented in [3].

DBpedia's ontology is based on classes (320 items) which form a subsumption hierarchy, the root element being owl:Thing, with a maximal depth of 5³. These classes are described by a total of 1650 different properties, forming a large set of RDF triples (580 million extracted from the English version of Wikipedia).

Even though DBpedia is now a worldwide project and provides pages in 125 languages, the English one is still the most represented. We can indeed find 4.58 million of things⁴ including 1,445,000 instances of the class *Person*, 735,000 places *Place*, 251,000 *Species* ... The number of instances described in this language is about three time larger than the second and third language (French and German).

As well as any RDF-structured dataset, DBpedia can be requested with SPARQL (which is an recursive acronym : SPARQL Protocol and RDF Query Language) queries. SPARQL allows the user to search, add, modify or delete RDF data available on the Internet, see [4] for more details about the language.

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 SELECT ?class
3 WHERE { <E> rdf:type ?class }
```

CODE 2.1: SPARQL Query : Search classes

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

¹<http://wiki.dbpedia.org/>

²<https://en.wikipedia.org/>

³Complete classes tree : <http://mappings.dbpedia.org/server/ontology/classes/>

⁴<http://wiki.dbpedia.org/about/facts-figures>

```
2 SELECT ?superClass
3 WHERE { <C> rdf:subClassOf ?superClass }
```

CODE 2.2: SPARQL Query : Search superclasses

Codes 2.1 and 2.2 present two simple and generic SPARQL search queries which return respectively the class(es) of a given entity E and the superclass(es) of a given class C using the *type* and *subClassOf* predicates.

DBpedia also provides useful web services and HTTP endpoints. DBpedia Spotlight, which highlights DBpedia concepts in an input text is described in [5] and further details about disambiguation using this service are presented in section 4.2. The official DBpedia SPARQL endpoint⁵ allows the user to send SPARQL queries to the online Virtuoso Triple Store by using the browser interface or by sending a HTTP request. We learn in [3] that the average amount of hits per day of this endpoint is of 2,910,410 for the 3.8 dataset version.

We find a lot of papers in the literature which mention DBpedia as an asset for systems based on the Semantic Web. If some of those papers are still in the research field, like [6] which proposes a music recommendation system built on top of DBpedia, others present “real” applications which are currently in use. We can cite for instance [7] which describes how the BBC⁶ uses DBpedia to backbone its publications.

2.3 Geonames

GeoNames⁷ is a geographical database which contains over 10 million geographical names. The most documented countries⁸ are the United States of America (2,203,094 names), Norway (600,008) and China (526,456). All resources are categorized into one out of nine classes and further subcategorized into one out of 645 codes⁹. Obviously,

⁵<http://dbpedia.org/sparql>

⁶<http://www.bbc.co.uk/>

⁷<http://www.geonames.org/>

⁸<http://www.geonames.org/statistics/>

⁹<http://www.geonames.org/export/codes.html>

the root element of Geonames' hierarchy is mother earth.

Like DBpedia, the Geonames' ontology makes possible the addition of new resources to the World Wide Web. However, Geonames distinguish the features' Concept from the RDF document about it. In consequence, each feature possesses two representation in Geonames. See the two following URIs as example :

URI1 <http://sws.geonames.org/8015555/>

URI2 <http://sws.geonames.org/8015555/about.rdf>

The first one stands for the "Notre Dame de Fourvière" church in Lyon, France. This URI is used when one wants to refers to the church itself. The second URI is the RDF document with what Geonames knows about the church, its latitude and longitude for instance, or some nearby locations.

In order to allow the user to browse its *Tremendous set of data* (Sir T. Berners-Lee), Geonames proposes a lot of REST-based web-services. In the case of image annotation, one in particular could be useful. Given that lot of recent numeric images contain EXIF data which embed GPS information such as the longitude and the latitude (see [8] for further details about the EXIF format), one could add geographic tags to the picture with the help of the *findNearbyPlaceName* service. Here is an example of usage :

Query : `findNearbyPlaceName?lat=48.566&lng=13.43&username=demo`

```

1 <geonames>
2   <geoname>
3     <toponymName>Passau</toponymName>
4     <name>Passau</name>
5     <lat>48.5665</lat>
6     <lng>13.43122</lng>
7     <geonameId>2855328</geonameId>
8     <countryCode>DE</countryCode>
9     <countryName>Germany</countryName>
10    <fcl>P</fcl>
11    <fcode>PPLA3</fcode>
12    <distance>0.00041</distance>
13  </geoname>

```

¹⁴ `</geonames>`

CODE 2.3: XML response

Given its specific domain, we found less use-cases of Geonames in the literature than DBpedia's. Some interesting papers have however been presented, like [9] which use several semantic resources including Geonames in order to detect named entities in "Agence France Presse" (AFP) wires.

2.4 WordNet

WordNet¹⁰ is a lexical database of English which has been presented for the first time in 1995 in [10]. It is hosted by the Princeton University, currently running version 3.1 but there are no current plan for a future release due to limited staffing.

Its structure is based on the concept of "synset" (synonym set), a set cognitive synonyms. WordNet distinguish among Types (common nouns, verbs. . .) and Instances (specific persons. . .). Synsets are interlinked using conceptual, semantic and lexical relations. The hierarchy is built by the use of the super-subordinate relation (or hyperonymy, hyponymy in WordNet's jargon). These relations implements the two directions of the "IS-A" expression. For instance, *fruit* is **hyperonym** of *apple* and *horse* is a **hyponym** of *animal*, the root element being "entity". Other relations are also provided, like the antonymy (opposite of synonymy) or the meronymy and its opposite holonymy which implements the "IS-PART-OF" relation : *finger* is a **meronym** of *hand*. All these relations are transitive.

This resource is useful if we are searching for entities. Since the maximal depth is of the ontology is of 16, the leafs are very detailed nouns (tsetse-fly, Yukon white birch, . . .) but it also contains more general concepts (vehicle, animal, . . .). WordNet contains at the moment 155,287 unique strings including 117,798 nouns.

¹⁰<https://wordnet.princeton.edu/>

It exists several ways to browse this resource. An online interface allows the user to manually query the dataset and to navigate in it through hyperlinks. For software and research purposes, the user has to download one of the released version of WordNet's dataset as well as a specific library according to the code language he's using.

Due to the lot of different relations between its synsets, WordNet has mostly been used for measuring semantic distances (see [11], [12] and [13]) or to disambiguate texts (in [14], [15] and [16]).

2.5 ImageNet

ImageNet¹¹ is an image database built on the WordNet hierarchy. It has been launch in 2009 and presented in [17]. Each of WordNet's synsets is depicted in ImageNet by a set of pictures (more than 500 in average). At the moment 14,197,122 images are referenced.

The stated aim of ImageNet is to serve as an asset for pedagogical and research purposes. It has for instance been used in [18] to measure the correlation between visual and semantic similarities or in a kindergarten in Canada to provide matching exercises to the children.

Despite the fact this resource is not as used as the previous ones, it was important to cite it in order to highlight that semantic concepts and images can be related even tough it's they are different kinds of medium.

¹¹<http://www.image-net.org/>

Chapter 3

Similarity Measures

This chapter aims to review several similarity measures. The first two sections describe semantic metrics when the last one shortly present a statistical way to measure concepts' relatedness. In order to illustrate those methods, we will use the figure 3.1 which is an handcrafted ontology.

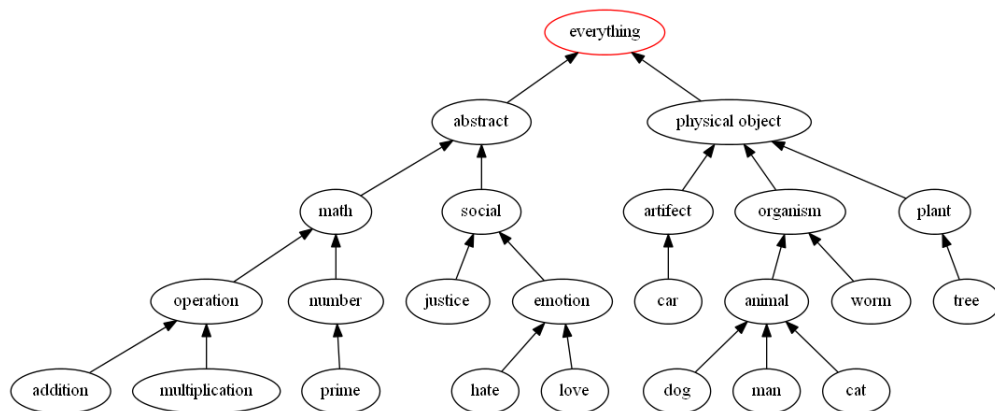


FIGURE 3.1: Sample ontology

3.1 Edge-based

3.1.1 Shortest Path

This measure is in fact a simple node-counting scheme (path). The similarity score is inversely proportional to the number of nodes along the shortest path between the

concepts. The shortest possible path occurs when the two concepts are the same, in which case the length is 1. Thus, the maximum similarity value is 1.

One has to be very careful with the use of this metric because it initially doesn't take into account neither the kind of relations nor their direction. Therefore, two concepts which are hierarchically related may be as similar as two leafs. Based on figure 3.1 we can for instance see that :

$$sim_{SP}(addition, social) = sim_{SP}(addition, prime) = 1/4 = 0.25 \quad (3.1)$$

It means that, with regards to the Shortest Path metric, addition and social are as similar as addition and prime.

3.1.2 Wu-Palmer

In [19] Wu and Palmer (1994) introduce a new scaled metric in order to translate English verbs into Mandarin Chinese. This measure takes into account the Least Common Subsummer (LCS)¹ of the 2 considered concepts. The formula is :

$$sim_{WP}(c_1, c_2) = 2 * depth(lcs(c_1, c_2)) / (depth(c_1) + depth(c_2)) \quad (3.2)$$

This means that $0 < sim_{WP} \leq 1$. The score can never be zero because the depth of the LCS is never zero since the depth of the root of a taxonomy is one. As for the shortest path, the score is one if the two input concepts are the same.

In our case, the results are the following :

$$sim_{WP}(addition, social) = 2 * 2 / (5 + 3) = 4/8 = 0.5 \quad (3.3)$$

$$sim_{WP}(addition, prime) = 2 * 3 / (5 + 5) = 6/10 = 0.6 \quad (3.4)$$

With *abstract* being the LCS of *addition* and *social* and *math* being the LCS of *addition* and *prime*. These results seems naively more normal than the one in equation (3.1).

¹Sometimes called LCA : Least Common Ancestor

3.1.3 Evolved Wu-Palmer

In [20], Zargayouna exposes a consequence of the Wu-Palmer formula which can, according to the context of use, gives false results. Indeed, in (3.2), the similarity between two concepts is related to their distance to the LCS. Therefore, the more “general” the LCS is, the less similar the concepts (and inversely). With this configuration is it possible to observe the following result with c_2 being a child of c_1 and c_3 a brother node.

$$sim_{WP}(c_1, c_2) < sim_{WP}(c_1, c_3) \quad (3.5)$$

Which can be an issue if one want to give an advantage to the father-child relation instead of the brother-brother one. Here is an example with our sample ontology from figure 3.1 :

$$sim_{WP}(addition, math) = 2 * 3 / (5 + 3) = 6/8 = 0.75 \quad (3.6)$$

$$sim_{WP}(addition, multiplication) = 2 * 4 / (5 + 5) = 8/10 = 0.8 \quad (3.7)$$

In order to achieve what’s mentioned above, the author create a virtual bottom node to which every leaf node is linked. Then, she integrates in the equation the maximum number of edges between this virtual node and the LCS as well as the number of edges between the LCS and each of the considered concepts.

$$sim_{WP++}(c_1, c_2) = 2 * depth(lcs(c_1, c_2)) / (depth(c_1) + depth(c_2) + spec(c_1, c_2)) \quad (3.8)$$

$$spec(c_1, c_2) = depth_{Bottom}(lcs(c_1, c_2)) * dist(LCS, c_1) * dist(LCS, c_2) \quad (3.9)$$

It is important to note that this modification of the Wu-Palmer formula rest on the fact that $dist(LCS, c)$ can be null if one of the concepts is the LCS which occurs in the father-child relation. In this case, the third part of the denominator becomes null and the similarity is the same as in (3.2). In the brother-brother relation yet, this part isn’t null and the similarity decrease. Let’s take the same examples :

$$sim_{WP++}(addition, math) = 2 * 3 / (5 + 3 + (3 * 0 * 2)) = 6/8 = 0.75 \quad (3.10)$$

$$sim_{WP++}(addition, multiplication) = 2 * 4 / (5 + 5 + (2 * 1 * 1)) = 8/12 = 0.66 \quad (3.11)$$

The results are those expected, and we can conclude that this measure works but that the user really has to know which behavior he needs before choosing WP or WP++.

3.2 Node-based

3.2.1 Resnik

In [21] Resnik presents a new measure of semantic similarity in an “IS-A” taxonomy, based on a notion he call “information content” (IC). Its work is established on the following hypothesis: the more the information two concepts share in common, the more similar they are. In order to evaluate the amount of information shared by the considered concepts Resnik use their LCS’ IC.

Given a chosen corpus, let $p(c)$ the probability to encountering an instance of c . Note that this probability is monotonic as one moves up in the taxonomy. It means that if c_1 IS-A c_2 , then $p(c_1) < p(c_2)$. For instance in our outology, $p(dog) < p(animal)$.

The Resnik’s formula stands as :

$$sim_R(c_1, c_2) = \log(p(LCS(c_1, c_2))) \quad (3.12)$$

We can not present results using our taxonomy because we would need a corpus to do so (like a book, composed of multiple sentences which contains words) but the original paper displays some in its Table 3.

3.2.2 Lin

Lin proposes in [22] a new measure based on Resnik’s (see 3.2.1). It adds to Resnik’s hypothesis the assumption that the similarity between c_1 and c_2 is also related to the differences between them; the more differences they have, the less similar they are.

Therefore, Lin modify the initial formula and adds a denominator :

$$sim_R(c_1, c_2) = \frac{\log(p(LCS(c_1, c_2)))}{\log(p(c_1)) + \log(p(c_2))} \quad (3.13)$$

3.2.3 Jiang and Conrath

Jiang and Conrath use the same hypotheses than Lin in [23], but their formula isn't a ratio but a difference :

$$sim_R(c_1, c_2) = \log(p(LCS(c_1, c_2))) - (\log(p(c_1)) + \log(p(c_2))) \quad (3.14)$$

3.3 Statistical similarity

It also exists statistic methods which evaluate the similarity of concepts. In [24], an algorithm is proposed based on the Jensen–Shannon divergence which measures the similarity between two probability distributions. It first calculates a probability distribution for each tag based on their co-occurrence with the most frequent words of the corpus and then compute the tags similarities.

This approach isn't semantic-based but it has been used as a comparison during our study.

Chapter 4

Disambiguation

Let's remember what we say about semantics in [2.1](#) : it is the study of meaning. The question that we will study here is the following : What happens when a word has several meanings ?

The disambiguation process aims to determine which one of the meanings is relevant in a specific context. In our case, we want to disambiguate the initial tags from a picture in order to propose relevant candidates. To achieve this, several approaches are presented with usages from the literature. Finally, we discuss the its interest in the image annotation process and the issues it may rise.

4.1 How does it work

4.2 Practical example : DBpedia Spotlight

Chapter 5

Existing services

5.1 Web service 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

5.2 Web service 2

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

5.3 Annotation via stats

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pelentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

Chapter 6

Conclusion

6.1 Section 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

Part II

Contribution

Chapter 7

Proposed Methodology

7.1 Section 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

Chapter 8

Proposed Architecture

8.1 Technology Choices

8.1.1 Java

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.1.2 Neo4j

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.1.3 NLP

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.1.4 DBpedia Spotlight

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.1.5 JAWS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.1.6 JENA

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie,

ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.1.7 JSoup

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.2 Graph Structure

8.2.1 Pro-Cons

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.2.2 Vertexes

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

8.2.3 Edges

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

Chapter 9

Experiments

9.1 Dataset

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

9.2 Code Explanation

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

9.3 Results and Analysis

9.3.1 Evaluation methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

9.3.2 Graph-based experiments

9.3.2.1 Direct Neighbors

9.3.2.2 Lists - WL

9.3.2.3 Lists - SL

9.3.3 Plain-text experiments

9.3.3.1 WikiLinks

9.3.3.2 WikiContent

Chapter 10

Conclusion

10.1 Section 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

- [1] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. 2001.
- [2] Eric Miller. An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19, 1998.
- [3] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al. Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 5:1–29, 2014.
- [4] Eric Prud’Hommeaux, Andy Seaborne, et al. Sparql query language for rdf. *W3C recommendation*, 15, 2008.
- [5] Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.
- [6] Alexandre Passant. dbrec—music recommendations using dbpedia. In *The Semantic Web–ISWC 2010*, pages 209–224. Springer, 2010.
- [7] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media meets semantic web—how the bbc uses dbpedia and linked data to make connections. In *The semantic web: research and applications*, pages 723–737. Springer, 2009.
- [8] Jelena Tešić. Metadata practices for consumer photos. *MultiMedia, IEEE*, 12(3): 86–92, 2005.
- [9] Rosa Stern and Benoît Sagot. Détection et résolution d’entités nommées dans des dépêches d’agence. In *Traitement Automatique des Langues Naturelles: TALN 2010*, 2010.

- [10] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [11] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2, pages 2–2, 2001.
- [12] Ray Richardson, A Smeaton, and John Murphy. Using wordnet as a knowledge base for measuring semantic similarity between words, 1994.
- [13] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics, 2009.
- [14] Philip Resnik. Disambiguating noun groupings with respect to wordnet senses. *arXiv preprint cmp-lg/9511006*, 1995.
- [15] Ellen M Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 171–180. ACM, 1993.
- [16] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer, 2002.
- [17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [18] Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1777–1784. IEEE, 2011.
- [19] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

- [20] Haïfa Zargayouna, Sylvie Salotti, et al. Mesure de similarité dans une ontologie pour l'indexation sémantique de documents xml. *IC 2004*, pages 249–260, 2004.
- [21] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.
- [22] Dekang Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998.
- [23] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.
- [24] Hatem Mousselly-Sergieh, Elöd Egyed-Zsigmond, G Gianini, M Döller, Harald Kosch, and Jean-Marie Pinon. Tag similarity in folksonomies. *INFORSID*, 2013.