Place in perspective: Extracting online information about Points of Interest

No Author Given

No Institute Given

Abstract. During the last few years, the amount of online descriptive information about places has reached reasonable dimensions for many cities in the world. Being such information mostly in Natural Language text, Information Extraction techniques are needed for obtaining the *meaning of places* that underlies these massive amounts of commonsense and user made sources. In this article, we show how we automatically label places using Information Extraction techniques applied to online resources such as Wikipedia, Yellow Pages and Yahoo!.

1 Introduction

In this paper, we present our approach to the challenge of assigning semantic annotations to places. These annotations are automatically extracted by applying web mining and information extraction techniques that have been thoroughly applied and tested in previous works[1]. In our case, we are particularly focused on extracting information that allows an interpreter to distinguish a place from other places that are spatially or conceptually close. In other words, the *meaning of a place* is a function of its most salient features, present in the textual descriptions found in online resources about that place. In our case, places correspond to *Points Of Interest* (POIs), as these are abundant in the web. By definition, a POI is a place with meaning to someone and, if it is available online, it is likely that its interest is shared by many people. In our approach, we first crawl the web to get a large quantity of POIs and then analyze each of them in order to obtain their individual *semantic index*: the set of words that best define it.

A system that is able to extract relevant semantics from places can be useful for any context aware system that behaves according to position. The level of information considered in this paper brings another layer to add to other sensors (GPS, accelerometer, compass, communications, etc.), eventually pushing forward the potential for intelligent behaviour. For example, a machine learning algorithm in a smartphone could be trained to present a different interface according to type of place (e.g. leisure, work, shopping). Other uses can be imagined, from navigation applications (e.g. navigating by concepts, searching for a place given related words) to analysis of social interactions and space use (e.g. finding correlations between POIs and presence of people). Beyond the scope of this article, we have also explored the dynamic information related to the places analysing the events happening in a given city.

We will start by giving the reader some essential background, and then we explain our methodology. We also present experiments and discuss a validation of the results. The present work is expected to be made publicly available by the expected date of publication of the article.

2 State Of the Art

2.1 Semantics of Place - From Space to Place

The difficulty in the unambiguous conceptualization of place comes along with its association to *space* and with the amount of different perspectives that may arise. Consider the simple question "Where am I?" and a sample of possible answers: relative to function ("I'm at work"); relative to someone ("I'm at my friend's place", "I'm with John"); relative to scale ("I'm in the US", "I'm in New York"; "I'm in 14th Street"); relative to objects ("I'm in my car", "I'm outside the stadium"). To this list of physical references, we can add the wealth of metaphoric creations of place (e.g. "I'm in second life", "My mind was somewhere else"). A place can be described with geographic, demographic, environmental, historical, and, perhaps also commercial attributes. The meaning of place derives from social conventions, their private or public nature, possibilities for communication, and many more [2,3]. Perhaps a simplification in this context, Harrison [4] works on the distinction between the concept of place from space, a place is generally a space with something added - social meaning, conventions, cultural understandings about role, function and nature. Often, it also has temporal properties; the same space can become different places at different times. Thus, a place exists once it has meaning for someone and the perception of this meaning is the main objective of our research.

As pointed by [5], absolute position such as the pair latitude/longitude is a poor representation of place. In our point of view, flexible representations that allow different perspectives become of greater importance, describing the world by commonsense and human-recognizable labels that best illustrate, in a synthetic way, the distinctive features of a given place contained in it (be it just a Point of Interest or a broad geographic Area).

2.2 Automatic Tagging - From text to terms

Lemmens and Deng [6] proposed a semi-automatic process of tag assignment which integrates knowledge from Semantic Web ontologies and the collection of Web2.0 tags. On a different direction, Rattenbury et al [7] identify places and events from tags that are assigned to photos on Flickr. They exploit the regularities on tags in which regards to time and space at several scales, so when "bursts" (sudden high intensities of a given tag in space or time) are found, they become an indicator of event or meaningful place. In the Web-a-Where project, Amitay et al [8] associate web pages to geographical locations to which they are related, also identifying the main "geographical focus". The "tag enrichment" process thus consists of finding words (normally Named Entities) that show potential for geo-referencing, and then applying a disambiguation taxonomy (e.g. “MA” with “Massachusetts” or “Haifa” with “Haifa/Israel/Asia”).

While our work focuses on the semantic aspect of location representation, we also take advantage of information available on the Web about public places. With the rapid growth of the World Wide Web, a continuously increasing number of commercial and non-commercial entities acquire presence on-line, whether through the deployment of proper web sites or by referral of related institutions. This presents an opportunity for identifying the information which describes how different people and communities relate to places, and by that enrich the representation of a Point Of Interest. Notwithstanding the effort of many, the Semantic Web is hardly becoming a reality, and, therefore, information is rarely structured or tagged with semantic meaning. Currently, it is widely accepted that the majority of on-line information contains unrestricted user-written text. Hence, we become dependent primarily on Information Extraction (IE) techniques for collecting and composing information on the Web.

Some relevant works can be referred in this realm, namely Open Calais [9] and Semantic Hacker [10], which focus on entity extraction from unstructured texts. They provide semantic indexes, although their focus is not restricted to information about space. A different approach, Scarlet [11], works on the extraction of the relations between concepts, an extremely challenging task within Information Extraction. Our approach shares references and methodologies with these works, but we emphasize information related to the place.

3 Automatic Labeling of a Point of Interest

3.1 KUSCO

Having a set of pages as input, Kusco [1] extracts a ranked list of concepts. This process includes Noun Phrase chunking and Named Entity Recognition (NER) using available NLP tools [12,13]. *Noun Phrase chunking* is made typically by partial (sometimes called ‘shallow’) parsers and extract clusters of words that represent people or objects. They tend to concentrate on identifying *base* noun phrases, which consist of a *head* noun, i.e., the main noun in the phrase, and its *left modifiers*, i.e, determiners and adjectives occurring just to the left of it. Named Entity Recognition tries to identify proper names in documents and may also classify these proper names as to whether they designate people, places, companies, organizations, and the like.

On completion of these subtasks, for each page, KUSCO ranks the concept with TF-IDF [14] (Term Frequency × Inverse Document Frequency) in order to extract the most relevant terms that will represent a given place. These nouns are contextualized on WordNet and thus can be thought not only as a word but more cognitively as a concept (specifically a synset - family of words having the same meaning, i.e., synonyms [15]). Given that each word in WordNet may have different meanings associated, its most frequent sense is selected to contextualize a given term. For example, the term "wine" has two meanings in WordNet: "ermented juice (of grapes especially)" or "a red as dark as red wine"; being the first meaning the most frequent used considering statistics from WordNet annotated corpus (Semcor[16]).

When using data from different sources, integration of information is imperative to avoid duplicates. To solve this problem we treat differently common nouns (generally denoting concepts) from proper nouns (generally Named Entities found). Although we use WordNet to find synonyms in the first group, we don't have a list of all possible entities in the world to match words from the second group. So, we take advantage of the relatively mature field of *String metrics* to find the distance between strings using an open-source available library with different algorithms implementations [17].

3.2 The perspectives

For any place, we build different lists of words, each list representing a "perspective" on the place, mostly dependent on the resource being analysed.

Open Web Perspective The *Open Web* perspective consists of crawling the web using a search engine (Yahoo!) given a POI name and address. The term "open" means that the search is not constrained to any particular web domain. The address is composed by the City name (where the POI is located) and is obtained from Gazetteers [1] available on Web. The search is made by the freely available Yahoo!Search API. We apply a heuristic that uses the geographical reference as another keyword in the search. Thus, assuming a POI is a tuple (Latitude, Longitude, Name), the final search query will be: <City Name> <Name>. To automatically select only pages centered on a given place, we filter out unuseful Web Pages with the following heuristics:(1) The title must contain the POI name; (2) The page body must contain an explicit reference to the POI geographical area; (3) Out of date pages will not be considered.

Wikipedia Perspective Wikipedia provides us with a massive database of partially structured textual information, currently about over 3 million topics. Plenty of relevant information about places is obtainable, both directly by searching for the actual Wikipedia page of a POI (e.g. Starbucks), and indirectly by finding information related to its category (e.g. Restaurant). We now present the two variations currently implemented, the red Wiki (indirect approach) and the yellow Wiki (direct approach).

Low-precision labeling: the Red Wiki

In the Red Wiki perspective, we extract the Wikipedia page corresponding to the identified category of a POI. Local POI directories are normally structured in a hierarchical tree of categories. This taxonomy may be created by the company

[1] A geographical dictionary generally including position and geographical names like Geonet Names Server and Geographic Names Information System [18].

itself or be collaboratively built by suggestion of users who feed the system with new POIs. We don't assume a rigid organization neither a consistent validation of such taxonomy. So, node duplication and multiple heritance may be a reality that a generic methodology must face. Actually, in our database it is normal for each POI to have multiple categories.

Since no API is currently available from those local directories studied, we have created a wrapper based on regular expressions in order to automatically extract the category taxonomy of each local directory. Only Yelp web site provides the complete list of categories, while Yahoo! Local only presents it through menu navigation along its web site. Curiously, this dynamics is also observed in the fact that this taxonomy is different depending on which city we are virtually visiting. Namely, Yahoo Local builds dynamically their menus, thus presenting proper taxonomies to distinct cities. Through time, this taxonomy grows with new types of services and places. In this way, by using specific wrappers to each POI provider, it is possible to run it periodically to integrate new categories in the respective stored taxonomy.

To contextualize each category in the corresponding Wikipedia article we base ourselves on string similarity between the category name and article title. We have opted for a top-down approach, from main categories to taxonomy leaves. To increase the confidence of this process, we manually disambiguate the main categories to start with and make sure that at least a more generic category will be connected to the Wikipages of its hypernym. When a POI has many categories, we obtain the articles for each one and consider the union of all the resulting articles as the source of analysis. Since there are many different combinations of categories, we can guarantee that each POI gets its own specific flavor of category analysis.

Medium-precision labeling: the Yellow Wiki

While the previous approach is centered on place category, here we focus our attention on Place name. We use string similarity to match Place name to Wikipage title in order to find the Wikipedia description for a given place. On a first glance, this method is efficient in mapping compound and rare place names such as 'Beth Israel Deaconess Medical Center' or 'Institute of Real State Management', however it can naively induce some wrong mappings for those places with very common names (e.g., Highway - a clothing accessories store in New York, Registry - a recruitment company in Boston, Energy Source - a batteries store in New York). We solved this problem by determining the specifcity of place names, and only considering those with high Information Content (IC)[19]. The Information Content of a concept is defined as the negative log likelihood, -logp(c), where p(c) is the probability of encountering such concept. For example, 'money' has less information content than 'nickel' as the probability of encountering the concept, p(Money), is larger than encountering the probability of p(Nickel) in a given corpus. For those names present in WordNet (e.g. Highway, Registry), IC is already calculated [16], while for those not present in WordNet, we heuristically assume that they are only considered by our approach if they are not a node in Wikipedia taxonomy, i.e., a Wikipage representing a Wikipedia category (case of Energy Source), but being only a Wikipedia article.

4 Experimental results

We have collected a large set of Points of Interest from Boston, New York and San Francisco. The extraction of words for those POIs (to what we call *enrichment*) needs some processing time. The average time for a POI analysis from the Open Web perspective is approximately 108 seconds, while Red and Yellow Wiki are 57 and 31 seconds, respectively. The Open Web is naturally more time consuming since it searches the entire web (using Yahoo! search engine), while any of the other perspectives uses a more bounded search. In Table 1, we present the overall statistics.

	New York	**Boston**	**San Francisco**	**Overall**
Yahoo!	183144	64133	94466	341743
YellowPages	7694	12878	-	20572
OpenWeb	757	2020	-	2777
Red Wiki	69011	20309	-	89320
Yellow Wiki	4400	1928	-	6328

Table 1. Above: number of POIs per perspective/city; Below: number of enriched POIs per perspective/city.

Regarding the words obtained, we have a total of 77558 different words, of which 9746 (12.6%) were also identified in WordNet. An analysis to these concepts was made regarding the average information content (IC) obtained. The IC [19] reflects the balance of specificity of the concept in a scale of 0 to 17. This average is 16.313395 (st.dev.=1.7263386), meaning that the concepts are in general very specific, thus carrying a rich content to the definition of POIs. This is however a risky game: if concepts are generic, the probability of being correct with respect to the place is much higher than when they are very specific. Since these words come from the actual text, in general they should be correct.

We show in Table 2 an excerpt of the “good” and “bad” examples found. This choice was made by the authors and intends to reveal the qualities and problems of the approach. A less subjective perspective on the results is presented in the next section. Except for the Red Wiki, we only put one category for each POI (many of them have more than one) to make the table clearer and let the reader understand the type of place.

The results from Open Web are extremely dependent on the initial search accuracy. In other words, if the correct webpage about the POI is found, then generally the results are acceptable, however this is not always simple to guarantee, depending on the nature of the POI. For example, if its name is a common noun (e.g. “Gap”), there will be too many unrelated pages, if it doesn’t have a

Name	**Categories**	**Terms**
Open Web		
Envirotech Incorporated	Waste and Environmental Consulting	Industrial Services, Asbestos Management, Mildew Removal, Asbestos Removal, Residential Services
Grasshopper	Telecommunications	Boston Telecommunications, Gary, Communication Services, Boston Business Directory, Telephone Communications
I Have A Dream Foundation	Educational Consulting	Boulder County, Dany Garcia, Arne Duncan, Jeffrey Gural, National Partners
Monroe Paint Distributors Incorporated	B2B Paint & Wall Coverings	movie theater, latitude, beauty salon, Delicious, Construction
Red Wiki		
Harvard Market	Grocery Stores	groceries, retailing, food, vegetables, products
Kim Depole Design Incorporated	Interior Design	office space, architects, private residence, code, decoration
Cambridge Library	Libraries	collection, library, information needs, public body, access points
Harvard Magazine	Marketing Agencies, News Services	pool, product, industry trade group, farmers, consumers
Yellow Wiki		
Boston Police Department	Law Enforcement	Massachusetts, law enforcement agency, correction, investigation, responsibility
TD Garden	Entertainment Venues	Boston Celtics, arena, Boston Blazers, naming rights, National Lacrosse League,
Starbucks Coffee	Coffee Houses	stores, Seattle, Washington, drip, Israeli
Blue Smoke	Steak Houses	Nora Roberts, Blue Smoke, Television film, novel

Table 2. Some examples from experiments (in each perspective, first 2 are "good" examples, last 2 "bad" examples).

webpage (e.g. it only exists in directory listings), there won't be any page. The Red Wiki perspective easily obtains meaningful words, although hardly specific to the POI, which is expectable since it works on its category. It is therefore a very "safe" perspective in terms of guaranteeing correctness of the obtained semantics. The Yellow Wiki can get much more refined results (e.g. the TD Garden is in fact the arena where Boston Celtics play NBA games) but it is more fragile when the wrong Wikipedia page is found (e.g. the Blue Smoke stake house is taken as a film with the same name) and is easily fooled by lateral information (e.g. Starbucks being from Seattle should be less relevant than for example for serving coffee or cappuccino).

5 Validation

We face an important challenge of understanding the actual quality of the results in terms of the *correctness* of the words assigned to places. The *ideal* list of words is by nature subjective. As referred above, a place can be defined according to different perspectives, and each perspective can vary with subject. In terms of validation, this raises difficult questions even for the typical user survey. A very large sample of people that *know* the specific places is necessary to achieve believable results, which then becomes unpractical and costly. We decided to analyze our results according to 2 dimensions: category consistency and coherence among perspectives. We also analyze the distribution of the words in each perspective.

Each POI has ultimately one category[2], so, in the *category consistency* validation, the task is to verify the stability of the word patterns according to those categories (15 for POIs). The first approach is to apply a clustering algorithm such as K-Means, where K corresponds to the number of different categories. After clustering with a training set, we apply a classification task: given the 5 top words of a POI from the test set, classify the POI in one of the categories. We apply 10 fold cross validation[3]. In order to get basic benchmarks to analyze the results, we set up two baselines: the *random baseline* consists of the accuracy of a random classifier (applied to all cases of the data set); the *fixed baseline* classifier selects the most popular class.

The resulting accuracy of clustering is in fact extremely poor, even when compared with the baselines. The highest value obtained (37.66%) was for the OpenWeb perspective which is actually lower than the *fixed baseline* of 47.49%(the accuracy obtained by a dumb model which basically always assigns the most popular category to any POI). This implies that either the word patterns are

[2] In reality, only a portion of the POIs have a single category, but we determined the *Least Common Subsummer* for POIs with multiple categories in the hierarchy, which consists of the most specific upper category that contains the categories of the POI in its descendants.

[3] Divide data set into 10 folds, each fold will become a test set to a model built with the remaining 9 folds[20].

not constant with respect to category or they are more elaborate than achievable with clustering algorithms. We tried Bayesian Networks, which are actually more common in text categorization, and the results improved considerably (accuracies from 57.12% to 97.3%). In Table 3, we summarize the results. The high value for the Red Wikipedia perspective reflects that our algorithm could extract sufficiently specific words from Wikipedia category definitions such that they become easily distinguishable form each other. This is interesting because many POIs (those that had multiple categories) were assigned a more generic category for classification, thus gathering in the same *class* POIs from different original areas of the category hierarchy and with many different words. Taking this into account, we can conclude that the original assignment of categories to the POIs is itself very consistent (e.g. Food & Dining subcategories are rarely mixed with Home & Garden ones).

For the Yellow Wikipedia perspective, the patterns are still extremely stable while, for Open Web , the results become less prominent. For Open Web a careful analysis reveals that there is still a reasonable quantity of noise in the indexes as we are collecting information from different Web sites with distinct templates and lateral information (e.g. advertisments, ads by Google, news headlines from RSS feeds) while in Wikipedia we have an available API to extract only useful and structured information. We can thus conclude that, for all perspectives, our system brings a degree of consistency that is relevant, particularly considering the two baselines.

	Rand. baseline	**Fixed baseline**	**K Means**	**Bayesian Network**
Red Wiki	9.199%	16.54%	24.40%	97.3%
Yellow Wiki	13.483%	23.25%	28.26%	66.56%
Open Web	23.781%	42.49%	18.71%	57.12%

Table 3. Category consistency results.

The analysis on coherence among perspectives allows us to see the stability of word patterns from the different sources. The assumption is that for the same POI the words from different perspectives should be related and/or similar. This relatedness is computed by Cosine Similarity [14] between indexes from different Perspectives for a given POI ranging from 0 (most dissimilar) to 1 (most similar). This analysis is however limited to the POIs that already have been analyzed for more than one perspective (which corresponds to less than 1000 overall). Firstly, we create a sparse matrix of Terms/Indexes ocurrences by weighting each ocurrence using TF-IDF weight of each term. Terms present in WordNet are contextualized as concepts (synsets) in order to identify synonyms from different indexes. For example, if we have the term “nightclub” on a given perspective mapped to WordNet (e.g.,“a spot that is open late at night and that provides entertainment (as singers or dancers) as well as dancing and food and drink”), any word representing this concept will be considered as a match (e.g. cabaret, night club, club, nightspot). The remaining terms are compared using string similarity in order to find little variation on names like "market intelligence" and "marketing intelligence". Table 4 presents computed similarity between perspective pairs of the same POI from the Boston Area. The Best Case of similarity is the closest pair to a given POI, and the Worst Case the farthest one. The overall Best and Worst cases are detailed in Table 5. Terms mapped into WordNet are complemented with synonyms enclosed by parentheses, "()". Looking at the examples, the dispersion between perspectives may be not a disadvantage but a richness acquired by the contribution of distinct terms from different perspectives. And, maybe, even for concepts not exactly synonymous but related (e.g, telecommunications, telephone) we can apply in the future other semantic similarity measures taking in account the meaning of each particular concept.

	#POIs	**Avg. Similarity**	**Best**	**Worst**
Open Web x Red Wiki	583	0.491+-0.362	0.996	0.008
Open Web x Yellow Wiki	52	0.368+-0.295	0.928	0.019
Red Wiki x Yellow Wiki	573	0.836+-0.122	0.990	0.288

Table 4. Analysis on coherence among perspectives.

Analysing these numbers, we can see that comparisons with Open Web perspective are less successful since their vocabulary is not bound by the Wikipedia domain. Furthermore, Red and Yellow Wiki perspectives are most coherent as they present the same related words in most cases. This is also confirmed by the highest average similarity, lowest standard deviation and reasonable size of sample set.

Dexter School	(categories: Parochial Schools, Elementary Schools, Middle Schools, High Schools, Preschools)
Perspective	**Terms**
Yellow Wiki	Grade(class, form, course), students, Schools, teamwork, Francis Caswell,...
Red Wiki	compulsory education, tuition(tuition fee), teachers, North America, students,...
Grasshopper (category: Telecommunications)	
Open Web	Boston Telecommunications, Gary, Communication Services, Boston Business Directory, Telephone Communications,...
Red Wiki	Telecomunication(telecom), modern times, telephone(phone, telephone set), inventors, semaphore,...

Table 5. An example of High similarity (above) and another of Low similarity (below)

We also randomly produced a sample of 420 Semantic Indexes (Red and Yellow Wiki) about Boston POIs which were manually validated by 28 volunteers who know the city in study, answering the question, for each word, whether it is related to the POI or not. We obtained a precision of 58% ($\sigma = 15\%$) and 56% ($\sigma = 20\%$) for Yellow and Red Wiki perspectives, respectively, considering all unanswered tags as invalid. In some cases even the volunteers disagree, reflecting the subjective nature of this information. Finally, we also check the shape of the word frequency histogram. In every perspective, we observed the distribution of words follows the typical long tail distribution that matches Zipf's law for word frequency [21], which was an expectable result.

6 Conclusions and Further Work

In this paper, we presented our work on semantic annotation of places from web resources. The implemented system could gather a massive amount of POIs and analyze a large portion of it, clearly enough for a valid analysis. The experiments show that the semantic indexes obtained have an average good quality, and we presented several different "perspectives" that can be used according to the context. When we need to guarantee the correctness of the words, we should use the Red Wiki perspective, which sacrifices the specificity of each POI for the analysis of its category (which brings normally correct results). When we are looking for exact information about a specific place, we can use the Yellow Wiki or the Open Web perspective. The former is preferable when the POI exists in the Wikipedia while the latter is the only option otherwise.

We believe that Location Based Services can improve their perception of Location Context through Semantic Enriched Places. Being able to infer implicit properties about places by semantic tags, they may relate places semantically closed that by classical representation (position, name, category) it would not be so cleary indentified. An interesting further step in our research consists of trying to deduce these semantic correlations and create a model of land use to understand the pattern of business/human occupation in some regions in the city.

To achieve a refinement of our system, we plan to implement a High-Precision place labeling for Wikipedia articles increasing the realiabity of Yellow Wiki, namely for specific names (with high Information Content) that are not related at all with the Place we analyze. For example, 'Apostrophe' is a store for commercial photographers in San Francisco, however searching in Wikipedia, we obtain the page describing the punctuation mark. This seems a gap in our methodology that could be solved by comparing the similarity between the place index with their corresponding category (Commercial Photographers). Another confidence level we can add to this mapping could be to verify if the geographic information contained in Wikipedia pages are related to a given place, be it by geo-reference annotations (presently available in more than 500000 articles [4]) or by textual patterns (country, city or neighborhood name).

References

1. Alves, A., Pereira, F.C., Biderman, A., Ratti, C.: Place enrichment by mining the web. In: Proc. of the Third European Conference on Ambient Intelligence. (2009)
2. Genereux, R., Ward, L., Russell, J.: The behavioral component in the meaning of places. Journal of Environmental Psychology **3** (1983) 43–55
3. Kramer, B.: Classification of generic places: Explorations with implications for evaluation. Journal of Environmental Psychology **15** (1995) 3–22
4. Harrison, S., Dourish, P.: Re-place-ing space: the roles of place and space in collaborative systems. In: CSCW '96, New York, NY, USA, ACM Press (1996) 67–76
5. Hightower, J.: From position to place. In: Proc. of the Workshop on Location-Aware Computing. (2003) 10–12 part of the Ubiquitous Computing Conference.
6. Lemmens, R., Deng, D.: Web 2.0 and semantic web: Clarifying the meaning of spatial features. In: Semantic Web meets Geopatial Applications. AGILE'08. (2008)
7. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: SIGIR, USA, ACM (2007) 103–110
8. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: geotagging web content. In: SIGIR '04, New York, NY, USA, ACM (2004) 273–280
9. Reuters, T.: Open calais api (2009)
10. TextWise: Semantic hacker api (2009)
11. Sabou, M., d'Aquin, M., Motta, E.: Exploring the semantic web as background knowledge for ontology matching. Journal of Data Semantics) (2008)
12. Ramshaw, L., Marcus, M.: Text Chunking using Transformation-Based Learning. In: WVLC, Cambridge, USA (1995)
13. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL '05. (2005) 363–370
14. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24**(5) (1988) 513–523
15. Fellbaum: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998)
16. Mihalcea, R.: Semcor semantically tagged corpus. Technical report, CiteSeer (1998)
17. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: IJCAI WS Inf. Integration. (2003) 73–78
18. Imagery, N., Agency, M.: Geonet names server (gns) (2007)
19. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: IJCAI. (1995) 448–453
20. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann (2005)
21. Zipf, G.K.: Selective Studies and the Principle of Relative Frequency in Language. Harvard University Press (1932)

[4] by Wikipedia Geoname Web Server http://www.geonames.org/export/Wikipedia-webservice.html