Acquiring Thesauri from Wikis by Exploiting Domain Models and Lexical Substitution

Claudio Giuliano¹, Alfio Massimiliano Gliozzo²,
Aldo Gangemi³, and Kateryna Tymoshenko⁴

^{1,4} FBK, Via Sommarive 18, Trento, Italy

^{2,3} STLab-CNR, Via Nomentana 56, 00161 Rome (RM) Italy
giuliano@fbk.eu, alfio.gliozzo@istc.cnr.it,
aldo.gangemi@istc.cnr.it, tymoshenko@fbk.eu

Abstract. Acquiring structured data from wikis is a problem of increasing interest in knowledge engineering and Semantic Web. In fact, collaboratively developed resources are growing in time, have high quality and are constantly updated. Among these problems, an area of interest is extracting thesauri from wikis. A thesaurus is a resource that lists words grouped together according to similarity of meaning, generally organized into sets of synonyms. Thesauri are useful for a large variety of applications, including information retrieval and knowledge engineering. Most information in wikis is expressed by means of natural language texts and internal links among Web pages, the so-called wikilinks. In this paper, an innovative method for inducing thesauri from Wikipedia is presented. It leverages on the Wikipedia structure to extract concepts and terms denoting them, obtaining a thesaurus that can be profitably used into applications. This method boosts sensibly precision and recall if applied to re-rank a state-of-the-art baseline approach. Finally, we discuss how to represent the extracted results in RDF/OWL, with respect to existing good practices.

1 Introduction

Acquiring structured knowledge from unstructured data is a crucial issue in the field of semantic technologies. Lexical knowledge plays a crucial role for applications dealing with natural language. Such knowledge, as it is usually processed by linguists, is typically represented following a structural paradigm in which *concepts are represented by a set of synonymous terms*. When one term refers to two different concepts, the term is ambiguous, while the fact that the same concept is expressed by two or more terms is called variability. Most computational lexicons are structured in this way, as variability and ambiguity are the more pervasive phenomena in language.

While defining concepts as sets of (nearly-) synonymous terms is a clear and pragmatic assumption among computational linguists, this is far less clear and pragmatic among domain experts, and even less among ontology engineers. However, the usefulness of lexicons like WordNet for ontology engineering techniques (ontology design, automatic mapping of ontologies) has eventually made that assumption acceptable, when revised appropriately. We will see how to reconcile the semantics of the results obtained through that assumption, with the formal semantics expected for ontologies.

Typically, the main weakness of computational lexicons is their lack of coverage when compared to actual usage of language. Lexical resources can be crafted by hand, however they typically suffer of poor coverage as all possible variants are typically not found by lexicographers. Corpus-based approaches can be used to enhance coverage, for example, to extract domain specific terminology or to find semantic relations among words. However, state-of-the-art approaches are not accurate enough, so domain experts are frequently involved in the process to provide additional validation.

The process of inducing terminology and paradigmatic relations among terms (e.g., synonymy and hyponymy) has been called *thesaurus induction* [6]. To this end, pattern-based approaches have been used [15], as well as distributional measures of lexical similarity based on word co-occurrence in documents [8] or in syntactic contexts [7]. Even though these approaches are scalable and cost-effective, their main limitation is that they can be used to discover relations among words, while semantic relations are defined between concepts. The term *thesaurus induction* is again a bit misleading when compared to what is known as a thesaurus in the Semantic Web literature. What computational linguists call thesauri are in fact a coarse approximation of mainstream thesauri known on the Semantic Web for the ability to be reengineered, e.g., through the SKOS vocabulary¹. A SKOS-compliant thesaurus is similar to linguists' thesauri in the sense that it is made of concepts, which have preferred and alternative labels. But realistic thesauri (that SKOS describes) are actually richer than that, and include a whole range of semantic relations between concepts: *broader*, *related*, *mapping*, etc. The results of (near-) synonymy, that thesaurus induction techniques show, should actually be matched against the richer model of thesauri. However, that classification detail cannot be easily achieved with purely statistical techniques, as precision hardly exceeds 40% with the best model.

In order to make thoroughly clear what are the potential benefits that might derive from bootstrapping thesauri, and formalizing them, we propose a novel technique to automatically acquiring thesauri from the analysis of the structure and content of wikis. Our work is contextualized in the area of research aimed at extracting knowledge bases from Wikipedia (see for example the DBPedia [1] and YAGO [21] projects).

In contrast to the previous corpus-based approaches, our technique is able to identify senses for each term, and then to extract (near-) synonyms specifically for each different sense. Another major limitation of previously proposed approaches is their limited lexical coverage. In fact, to measure the distributional similarity between two words, it is necessary to collect multiple occurrences of those words in a given corpus to define reasonably good context representation. This is particularly problematic when dealing with domain specific terms, whose occurrences are somewhat rare. In contrast, our method is scalable and ensures high coverage, as it relies on two large-scale resources: Wikipedia and the Web 1T 5-gram corpus [2].

Our technique is based on two fundamental linguistic properties characterizing paradigmatic relations between words: the *domain restriction hypothesis* and the *lexical substitutability*. The former states that paradigmatic relations (e.g., synonymy and hyperonymy) holds mainly among terms belonging to the same domain, where domains represent regions in the language characterized by topical similarity and, in our case,

¹ <http://www.w3.org/TR/skos-reference/skos.rdf>

modeled by latent semantic analysis (LSA). The latter claims that terms are synonyms if they can be substituted in a particular context preserving the original meaning. Relying on these assumptions, we adapted the lexical substitution technique proposed by [11] to the problem of finding synonymous terms for terms in context. For example, the term *book* in the sentence “All Gold Canyon is a book written by Jack London” can be substituted with *novel* preserving the original meaning. It means that *book*, in this particular sense, is a (near-) synonym of *novel*.

Another innovation of our approach is that we enhance the thesaurus induction process by relying on the existing conceptual structure provided by Wikipedia, assuming that each article identifies a different concept that can be expressed by one or more terms. Our goal is then to discover sets of (near-) synonymous terms describing that concept. To this end, we extract the set of candidate (near-) synonyms from the Wikipedia article and its incoming wiki links. Then, we choose all those terms belonging to the domain of the article by applying LSA similarity. Finally, we rank each term by applying the lexical substitution technique. The result is a set of synonyms that can be used as a thesaurus entry for the concept.

We apply this method to detect all possible synonyms for subset of the most frequently visited articles in Wikipedia, and we evaluated our approach by comparing the results with respect to the corresponding synonym lists provided by the Oxford American Writers Thesaurus, after an handcrafted mapping of these articles to the corresponding dictionary senses. Experiments show that our method is accurate and improves the baseline approach.

The rest of the paper is structured as follows. In Section 2, we identify the analogies between the structures of wikis and thesauri. Section 3 describes the two basic building blocks of our algorithm for detecting paradigmatic relations: semantic domains and lexical substitutability. Section 4 reports the experimental results, Section 5 describes a reengineering pattern to formalize linguistic thesauri in OWL-RDF, Section 6 presents related work, and Section 7 concludes the paper suggesting some interesting applications of our method that we are going to explore in the future.

2 Deriving Conceptual Structures from Wikis

The wiki technology is changing the Web, leading to the success of Web 2.0 applications, such as Wikipedia. Wikis are currently adopted in many applications, including corporate knowledge management systems, collaborative development of lexical and encyclopedic resources, social networks, e-learning systems, and so on. The attractive feature of wikis is that their users are implicitly asked to collaboratively create a shared conceptual space, where each concept is described by a Web page mostly composed by highly focused natural language text concerning that concept. When a new concept of interest for the domain of the wiki is identified, the user creates an “empty” page where other users or her/himself can explain it in detail. On the other hand, when the concept of interest is already in the wiki, the user can link to it. The anchor text of the link may be a synonym or an equivalent expression of the term adopted to name the target page describing the concept.

Therefore, the wiki can be used as a basis to derive a network of relations between concepts (the Web pages) and terms (anchor texts of the internal links present in other wiki pages). More formally, we can describe the conceptual structure derived by a wiki as follows. Let C be the set of concepts, and T be the vocabulary of terms. From each wiki page we derive a concept, while terms, which are anchor texts of links to the page, become elements of T . Each wikilink describes a relation $M(t, c)$ between a term and a concept. For a given term t_i , the set of its possible senses is $S(t_i) = \{c_j \in C | M(t_i, c_j)\}$. For a given concept c_i , the set of its possible lexicalizations is returned by $L(c_i) = \{t_j \in M(t_j, c_i)\}$. The number of links connecting a given term to a given concept can be used to infer a prior probability for the term t_i having sense $c_j \in S(t_i)$, by adopting the Formula 1.

$$P(c_j | t_i) = \frac{|M(t_i, c_j)|}{|M(t_i, c_z) | c_z \in S(t_i)|} \quad (1)$$

The structure described above is a thesaurus, as synonyms are represented and connected to concepts. In fact, [19] and [16] followed this approach to extract a thesaurus out of the wikilink structure.

On the other hand, wikis are collaboratively developed resources, most of the time composed very quickly by non-professional users. As a consequence, errors and bad practices are very frequent. For example, the article describing the concept *book* is linked to by the following terms: *monograph*, *manual*, *work*, *novel*, *booklet*, *father*, *guide*, *literary work*, *textbook*, *folio*, *books*, *treatise*, *print*, and *bookshop*, *book*. But, for example, only *manual*, *novel*, and *treatise* are listed as synonyms in the Oxford American Writers Thesaurus. The list obtained from wiki structure supplies new synonyms like *textbook*, but it introduces obvious errors like *father* as well.

Therefore, the use of the Wikipedia structure is not a viable solution for inducing high quality thesauri, even if it provides a rich starting point. In particular, we identified the following limitations.

- Typographical mistakes in the anchor texts. E.g., *politik* linked to *wiki/Politics*.
- Wrong links assigned to terms. E.g. terms *Article* and *Money* were linked by someone to the wiki page *wiki/Philosophy*.
- Bad selection of the term to link to the wiki page. E.g., *where prohibited* was made an anchor text for *wiki/Book*.
- Coreferences linked to concepts instead of terms. E.g., *this guy* linked to *wiki/Pope*.
- Terms having different syntactic roles are actually referred to the same page. For example, *political* is an anchor for the page *wiki/Politics*.
- Many possible lexical variants of concepts are missing, as users typically do a search before deciding whether to create a new wiki page or linking to an existing one. E.g., *herb* is missing among the lexical variants of the concept *Plant*.

In the following sections, we propose a method relying on semantic domains and lexical substitution to solve all the above problems.

Table 1. Ranked lists of possible synonymous terms for the concept *book* extracted adopting different criteria (from top to bottom): Wiki links, ranked by frequency (baseline), nouns in the article text, ranked by frequency, union of wikilinks and terms ranked by LSA, and by the combined method exploiting lexical substitution and LSA

Method	(Near-) Synonyms
Wiki Links	book, monograph, manual, work, novel, booklet, father, guide, literary work, textbook, folio, books, treatise, print, bookshop, book '
Page Terms	ANSI, Abaca, Ages, . . . , E-book, book, bookbinder, . . . , novel, novels, number, numbers, . . . , purposes, queen, quo, quotation, . . . , stem, stems, stock, stories, story, . . . , values, varies, vellum, vergilius, version, vine, vines, virgil, vitriol, vivarium, volume, word, . . . , year, young, yun-ui
LSA	Book, book, Books, books, story, Reading, reading, Publication, publication, reader, Author, author, stories, readers, Publishers, publishers, Novel, novel, reprint, work, . . .
Lexical Subst.	book, story, magazine, novel, publication, chapter, writing, new, people, print, work, literature, reading, time, author, young, history, edition, collection, character, link, page, place, reference, english, press, review, note, account, kind, publisher, . . .

3 Discovering Synonymy

We aim at capturing two characteristic properties of paradigmatic relations for synonymy detection:

- *Domain properties*: if a semantic relation between two terms X and Y holds, both X and Y tend to belong to the same semantic domain [14].
- *Lexical substitutability*: if two terms X and Y are paradigmatically related, they can be mutually substituted in text preserving the meaning.

In the following two subsection we describe each of them and we show their exploitation for thesauri induction.

3.1 The Domain Restriction Hypothesis

Semantic domains are at the basis of lexical semantics. Domain properties, also known as domain restriction hypothesis, are illustrated in Figure 1, where the probability for two terms to be related in WordNet [9] by a paradigmatic relation is contrasted to their domain similarity, measured by computing the cosine similarity between their corresponding vectors in the LSA space.

The main advantage of adopting semantic domains for thesauri extraction is that they allow us to impose a domain restriction on the set of candidate pairs of related terms. In fact, semantic relations can be established mainly among terms in the same semantic domain, while terms belonging to different fields are mostly unrelated.

Domain similarity between terms can be captured by measuring the LSA similarity between them. For our purposes, we defined a domain model by unsupervised learning on the Wikipedia text, following the procedure described in [13]. c_i and t_j are defined in the LSA space, estimated by the Singular Valued Decomposition for the term by

Fig. 1. Probability of finding paradigmatic relations in WordNet contrasted to domain similarity estimated in the British National Corpus [14]

document matrix extracted from the subset of the 500,000 most visited pages of the English Wikipedia. As a result, the similarity between a concept c_i , described by the corresponding Wikipedia article, and a term t_j , can be estimated by $\cos(c_i, t_j)$ ².

Domain similarity is then used to identify sets of candidate terms to fill the thesaurus entry representing each concept c . We first collect a set of the candidate lexicalizations $L(c) \subset T$ for a target article (concept) c by merging the results of the following two strategies:

1. Selecting all nouns extracted from the analysis of the text in the Wikipedia article associated to c . For example the second row of Table 1 shows the list of terms obtained after the analysis of the article corresponding to the concept *book*.
2. Selecting words used to link the target page from other articles, i.e. the lexicalizations $L(c)$ extracted by the wiki structure described in Section 2. For example, the first row of Table 1 shows the list of terms obtained from the wikilinks pointing to the article *book*. We adopted this strategy to define our baseline.

Then, we ranked all terms in the set $t_i \in L(c)$ according to their LSA similarity with respect to the target concept $\cos(c, t_j)$, and we filtered out all those terms having similarity below a threshold, that we fixed to 0.5 in all experiments described in this paper. Using such threshold we were able to filter out up to 80% of the terms in the set. This step is crucial to boost the efficiency of the synonymy induction step, as the final ranking, based on lexical substitution, is computationally intensive and constitute a bottleneck for the overall efficiency of our method.

3.2 Finding Synonyms by Lexical Substitution

As a second step, the candidate terms having high domain similarity with respect to the target concept are re-ranked by lexical substitution. The lexical substitution is a textual

² The SVD process exploited 400 dimensions at unsupervised learning time and took less than 2h on a dual core processor exploiting 4GB of RAM.

entailment subtask in which the system is asked to provide one or more terms that can be substituted to w in a particular context $H_w = H^l w H^r$, generating a sentence $H_e = H^l e H^r$ such that both $H_w \rightarrow H_e$ and $H_e \rightarrow H_w$ hold, where H^l and H^r denote the left and the right context of w , respectively.

[17] recently proposed this task in the context of the SemEval-2007 evaluation campaign. The state-of-the-art system is quite accurate and totally unsupervised [12], as it only exploits an existing dictionary and a huge language model. For each target word, a set of possible candidate words is extracted by the dictionary, and then substituted to the target word generating new sentences. Then, the “plausibility” of each sentence is measured by looking for the frequency of the generated sentences in the Web 1T 5-gram corpus, and the candidate words are then ranked according to that plausibility score.

In order to apply this technique to our problem we cannot rely on the availability of existing dictionaries containing candidate terms, as our goal is to discover such terms. In addition, we are not interested in substituting terms into a single occurrence of a word in context, but we are rather interested in the combined results coming from the substitution of candidate words into a large set of occurrences, i.e., all wiki links pointing to a particular article.

For a given concept c and a set of its candidate lexicalizations $L(c)$, our substitution algorithm works as follows:

Step 1. For each concept c , we collect all the sentences S containing a wiki link i pointing to the Wikipedia article associated to c . For instance, a sentence containing a link to the concept *book* is “The most printed *book* in history”

Step 2. Then, for each sentence $s \in S$, we derive a hypothesis phrase by replacing the link i with a candidate synonym $j \in L(c)$. For instance, from the previous example, we derive “The most printed *novel* in history”.

Step 3. For each hypothesis phrase h_j , we calculate a plausibility score v_j using a variant of the scoring procedure defined in [11]. In our case, v_j is given by the sum of the pointwise mutual information (PMI) of all the n -grams ($1 < n \leq 5$) that contain j divided by the self-information (SI) of the right and left contexts. PMI and SI are defined by the Equations 2 and 3, respectively.

$$PMI = \log \frac{p(t_1, t_2 \dots t_l)}{p(t_1)p(t_2) \dots p(t_l)} = v_j \quad (2)$$

$$SI = -\log p(t_1, t_2 \dots t_l) = v_j \quad (3)$$

Dividing by the self-information allows us to penalize the hypotheses that have contexts with a low information content, such as sequences of stop words. The frequency of the n -grams is estimated from the Web 1T 5-gram corpus. For instance, from the hypothesis phrase “The most printed *novel* in history”, we generate and score 10 n -grams, some of them are: “The most printed *novel* in”, “printed *novel* in history”, “*novel* in history”

Step 4. Finally, to obtain an overall score v_t for the term t , we sum the scores obtained by the substitution in all sentences, as defined in Equation 4.

$$v_t = \sum_{l=1}^N v_l, \quad (4)$$

where N is the number of sentences containing a wiki link to the target article c .

We used the this procedure to rank all candidate terms selected in the domain restriction phase. The top ranked terms for each concept provide a different thesaurus entry, associated to the corresponding Wikipedia article.

4 Evaluation

For the evaluation, we selected the target concepts among the most visited articles in Wikipedia, and we automatically extracted thesaurus entries for each of them by contrasting our method with a baseline, provided by a replication of the method described in [19].

The list of concepts evaluated so far is the following: *Sex, Pornography, Love, Politics, Book, Earth, Map, Computer, Cat, Game, Dictionary, Ejaculation, Lesbian, Sport, Cancer, God, Virus, Animal, Heroin, Horse, Film, Human, Condom, Snake, Flower, Evolution, Beer, Statistics, Religion, Communication, Management, Insomnia, Erection, Death, Earthquake, People, Insurance, Graffiti, Research, Puberty, Cannabis, Marriage, Gun, Socialism, Narcissism, Twilight, Sleep, Architecture, Stroke, Leaf.*

4.1 Gold Standard

We handcrafted a gold standard thesaurus entry by mapping each concept into the appropriate sense of the Oxford American Writers Thesaurus. For simplicity, we did not consider multi-words and parts of speech other than nouns. Below, we report a sample of the gold standard: the first word is the name of the Wikipedia article, and the following words are its synonyms extracted from the appropriate sense in the dictionary.

book hardback, storybook, volume, treatise, publication, e-book, manual, novel, title, tome, paperback, anthology
earth planet, world, globe
computer laptop, terminal, mainframe, pc, desktop
dictionary wordbook, lexicon, glossary, thesaurus
cancer lymphoma, sarcoma, melanoma, malignancy, myeloma, tumor, carcinoma
virus contagion, disease, bug, infection

The output of our thesaurus induction method is a ranked list of terms for each concept, that we compared with the gold standard, reporting precision and recall figures.

4.2 Baseline

As a baseline method, we implemented the heuristic described in Section 2. We extracted all terms contained by wikilinks pointing to each article, and we ranked them by frequency (i.e., the number of times they have been used to link to the target article). For each term, we computed the precision at different levels of recall, following an information retrieval paradigm.

The micro-average figures, reported in Figure 2, show that this method is not very accurate. In fact, the maximum precision is close to 0.2 at very low recall. The break-even point is 0.18.

Fig. 2. Precision/Recall curves for the baseline and our system

4.3 Advanced Methods

We built the domain model from the 200,000 most visited Wikipedia articles. After removing terms that occur less than 5 times, the resulting dictionaries contain about 300,000 terms. We used the SVDLIBC package³ to compute the singular value decomposition, truncated to 100 dimensions. All the experiments were performed using the LIBSVM package [5].

Then, for each article corresponding to a concept in the gold standard, we adopted the bag of word vector described above to provide a list of candidate terms, that we merge to the set of words returned by the baseline method. Then, we filtered out from this set all those terms having a cosine similarity with the target article’s vector below 0.5. This threshold allows us to drastically reduce the number of candidates but preserving with high probability the (near-) synonyms. The third row of Table 1 exemplifies the list of terms obtained so far for the concept *book*.

Then we collected all sentences containing wiki links to the target article, and we generated new sentences by substituting all candidate terms collected above. Finally, we computed a plausibility score for each of those sentences, on the basis of which we derived the final score for each candidate synonymous term by adopting Formula 4. The resulting sets are then ranked by applying the lexical substitution algorithm described in Section 3.2.

Results are reported in Figure 2, we compare the two methods by means of the precision/recall curves. The lexical substitution method combined with the domain restriction hypothesis significantly boosts the quality of the baseline..

4.4 Qualitative Evaluation

From a qualitative analysis of the results, it seems that most terms that do not match the gold standard entry are not actually weird errors, but rather they follow under the definition of (near-) synonymy. For example, looking at the fourth row of Table 1 and

³ <http://tedlab.mit.edu/~dr/svdlbc/>

comparing it to the gold standard reported in Section 4.1, many high quality terms such as print, monograph, booklet, folio, guide, textbook, treatise, do not match the gold standard. However, they include hyperonymy, hyponymy and co-hyponymy terms. Such terms are very useful in developing computational thesauri, especially as far as applications like semantic information retrieval are the final goal. For example, the terms acquired so far can be used to improve search in Wikipedia itself, or to expand the lexical part of DBpedia, bridging knowledge to language in a much smarter way.

5 RDF-izing Linguistically-Induced Thesauri

A linguistically-induced thesaurus is currently understood as a “poor” thesaurus, which only contains concepts and (preferred or alternate) labels. The linguistic semantics, by which a concept is equivalent to a set of near-synonymous terms can be represented in different ways, and we show here three possible solutions: one based on SKOS, one on WordNet OWL, and a hybrid one.

However, before exemplifying the solutions, we should note that the assumption of near-synonymy between the terms (found as statistically meaningful), associated with a concept (i.e. the entities Wikipedia pages are about) can be counter-intuitive for lexical semantics.

Near-synonymy is expected to catch a special kind of similarity between terms, which makes them *substitutable* in a certain class of contexts, modulo certain local constraints. For example, *book* can be substituted with *volume* in many generic contexts (referring to physical objects), or with *script* in more specialistic contexts (drama playing), but a lexicographer would hardly make it substitutable with *magazine* or *writing*: the last two would be considered similar in a different sense (i.e. hyponymy, overlap, or association).

The notion of substitutability assumed in thesaurus induction is broader than the one assumed in lexical semantics, and the representations we present have to be adjusted to this broader notion, which also covers what lexicographers would classify as e.g. hyponymy or association.

The first candidate representation consists in using the relation between a skos:Concept and its skos:label(s): this representation basically uses literals as values for the skos:label property, and therefore has problems when we need to provide attributes to the terms. In a new extension of SKOS⁴, a label is represented as an individual from the class skosxl:Label, with which literal values can be associated. This gives also room for representing a large set of data about preferred or alternative labels: creators, dates, provenance, etc.

For example, the relation between: **Book** and {book, novel, magazine, textbook, publication, writing} can be represented as a set of RDF triples:

```

:Book a skos:Concept ; skosxl:prefLabel :book ;
      skosxl:altLabel :novel ; skosxl:altLabel :magazine ;
      skosxl:altLabel :textbook ; skosxl:altLabel :writing ;
      skosxl:altLabel :publication

```

with :book a skosxl:Label, etc.

⁴ <http://www.w3.org/TR/skos-reference/skos-xl.rdf>

Fig. 3. Precision/Recall curves for the baseline and our system used to re-rank the baseline

With this representation, the semantics of thesaurus induction (concepts against relevant associated terms) is represented, since `skos:Concept` and `skosxl:Label` are disjoint classes, but it is confusing for a user to find “labels” of e.g. the concept *Book* that include *novel*, *magazine*, etc., since most people is accustomed to the lexicographic near-synonymy assumption. The intuition for an English speaker would possibly be that the concept *Book* has more specific senses like *novel* and *textbook*, more generic ones like *publication*, overlapping terms like *magazine*, and related senses like *writing*. The most correct representation would be therefore to use the generic `skos:semanticRelation` between concepts, which can be eventually specialized as broader, related, etc. after automatic or manual evolution. In this case, either concepts or terms from thesaurus induction should be classified as `skos:Concept(s)`.

However, in this way we are not getting much to the point, since in order to preserve the lexicographic intuition, we lose the original thesaurus induction semantics.

An alternative to using SKOS is to reuse a meta-model that is closer to computational linguists’ intuition, i.e. the WordNet OWL vocabulary⁵. In that case, the class `wordnet:Synset` has the intended meaning of a set of near-synonymous word senses, which are linked through the `wordnet:containsWordSense` relation. A `wordnet:WordSense` is on its turn linked to words through the `wordnet:word` relation to `wordnet:Word`. This representation seems better, since it allows to representing terms either as word senses or words, and still retains the possibility to link them to synsets (concepts). However, even here the exceptions to intuitive near-synonymy mentioned above are still valid: some relations look closer to `wordnet:hyponymOf`, others to `wordnet:hypernymOf`, others do not have any correspondence at all. The relation `wordnet:containsWordSense` has a meaning that conflicts with e.g. `wordnet:hyponymOf`, and in WordNet no relation represents e.g. the one between *Book* and *writing*.

⁵ <http://www.w3.org/2001/sw/BestPractices/WNET/tf.html>

A third alternative consists in building an ad-hoc vocabulary for thesaurus induction semantics. This solution can be implemented by merging the positive aspects of both SKOS and OWL-WordNet: concepts are represented as `wordnet:Synsets`, terms as `wordnet:Words` mapped to `skosxl:Label(s)`, while new individuals for each word sense of a word can be created, and linked to their synset through the `skos:semanticRelation` relation, for example:

```

                                :Book a wordnet:Synset ;
    skos:semanticRelation :wordsense-book ;
    skos:semanticRelation :wordsense-novel ;
    skos:semanticRelation :wordsense-magazine ;
    skos:semanticRelation :wordsense-textbook ;
    skos:semanticRelation :wordsense-publication ;
    skos:semanticRelation :wordsense-writing

```

with `:wordsense-novel a wordnet:WordSense` ; and `wordnet:word:novel` which is both a `wordnet:Word`, a `skosxl:Label`, etc.

The third hybrid representation can also be implemented by designing a vocabulary that catches exactly the thesaurus induction semantics, and then by mapping the vocabulary to SKOS and OWL-WordNet.

6 Related Work on Thesaurus Construction

Among the approaches to semantic relatedness evaluation and thesauri induction there are the pattern-based approaches [15], as well as distributional measures of lexical similarity based on word co-occurrence in documents [8] or in syntactic contexts [7]. In [22] a thesaurus is generated by usage of a syntactically constrained Vector Space Model (VSM). In this approach each dependency from a set of predefined syntactic dependencies is represented by different subsets of VSM. Sets of words related to a given source word are found distinctly in each of the VSM and then overlapped.

Many of the latest approaches to the task are based on usage of Wikipedia as the source of background knowledge. Among the first were [20], who exploited the Wikipedia pages texts and category structure to calculate text overlap, taxonomy and information based relatedness measures. Given two words **i** and **j** they retrieved the corresponding Wikipedia pages (or articles), to content of which text overlap similarity based measures could be applied. The pages were retrieved just by querying a wiki page named **i** or **j**. In the case when one or both retrieved pages are disambiguation pages, disambiguation heuristic was applied. Exploiting sets of categories C_i and C_j , to which words belong, [20] obtained least common subsumer for each category pair to be able to calculate information based measures. Finally, sets of paths between the C_i and C_j pairs gave input to taxonomy similarity measures. The correlation with human relatedness evaluation is 0.19-0.48.

[10] proposed Explicit Semantic Analysis (ESA) for measuring semantic relatedness of both words and texts. They build the semantic interpreter, which maps input text/words into a weighted vector of relevant Wikipedia concepts. The relevant

Wikipedia concepts are titles of Wikipedia pages, which contain the words/texts. The semantic relatedness of two texts/words is computed then applying cosine metric to corresponding weighted concept vectors. The correlation with human relatedness evaluation is 0.75.

[18] measure semantic relatedness between the concepts using Wikipedia link structure. Given two terms, they obtain all wikipedia articles to which these terms may refer and compute semantic relatedness between all possible pairs of articles combining two relatedness measures. First one is similar to *tfidf*, but uses wikipedia links and pages instead of terms and documents. The second is based on the Normalized Google Distance, which takes into account term co-occurrences of terms on the web-pages, where links and wikipedia articles are used instead of terms and web-pages correspondingly. The terms relatedness is then evaluated as a combination of the highest relatedness achieved among pairs of articles with term co-occurrence frequency in wikipedia anchor texts. The correlation with human relatedness evaluation for words is 0.78.

[19] and [16] propose to use Wikipedia directly as a building material for an association thesaurus. Therefore, they do not encounter the disambiguation problem. In [19], the thesaurus is constructed using *pfibf* (path frequency inversed backward frequency), which calculates the relativity between two concepts I and J represented by two articles V_i and V_j . *pfibf* exploits information about the number of paths from V_i to V_j , the lengths of these paths and the number of backward links to articles. The latter is used to penalize popular articles, which are highly related to a big number of articles. The concept precision among 10-30 top terms varies within limits 66.7-85.9. In the later work ([16]) they propose a much faster associative thesaurus construction method based on link co-occurrence analysis. The links are considered to be concepts, two links are said to be the same even if they have different anchor texts. In order to calculate relatedness between two links each link is represented as a combined vector (*cv*). *Cv* is a linear combination of a *link* and *tfidf* vector. *Link* vector is a vector, whose positions are first-order co-occurrences of the given link with other links in the Wikipedia. The first-order co-occurrence of two links is computed using the co-occurrence frequencies of these links with all other links over Wikipedia. Accuracy measure is 0.59-0.69.

In addition, different techniques for acquiring terms, synonyms and semantic relations have been presented in the field of ontology learning [4,3].

7 Conclusion and Future Work

In this paper, we presented an innovative method for inducing thesauri from Wikipedia. Our method leverage on the Wikipedia structure to find out target concepts for the thesaurus, which is in itself an innovation if compared to traditional method relying on words only. In addition to that, we applied an innovative method for detecting possible lexicalizations of those concepts, showing that both precision and recall, in this settings corresponding to the quality of the thesaurus entry and to its lexical coverage, can be significantly boosted. A qualitative assessment of the results shows that our method is actually able to find (near-) synonymous terms that can be used to automatically populate thesaurus entries if the strict requirement of synonymy is relaxed.

For the future, we are going to perform an extensive analysis of the full Wikipedia aimed at extracting lexical expressions for each concept represented by a Wikipedia

page. In fact, our method is largely scalable: with the present (prototypical) implementation we were able to process each page in few seconds by using a standard PC. In addition, we are going to extract thesauri for many languages in Wikipedia, as our method is totally language independent and only requires part of speech tagging of texts.

Exploiting such a large scale thesaurus into applications seems very promising, as the main weakness of knowledge based approach for retrieval is very often coverage. For example, the extracted synonyms for each page can be used for improving the index of Wikipedia for (cross lingual) information retrieval.

Last, but not least, our technique provides an explanatory application of two fundamental properties of lexical semantics, the domain restriction hypothesis and the lexical substitutability principle, showing that a combination of them can be almost always used to figure out solutions to problems involving lexical semantics, providing yet another experimental proof of their universal validity.

Acknowledgments

Claudio Giuliano is supported by the X-Media project (<http://www.x-media-project.org>), sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978 and the ITCH project (<http://itch.fbk.eu>), sponsored by the Italian Ministry of University and Research and by the Autonomous Province of Trento. Aldo Gangemi is supported by the NeOn project (<http://www.neon-project.org>) and the IKS project (<http://www.iks-project.eu>), sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-2005-027595 and IST 231527, respectively. Kateryna Timoshenko is supported by the ITCH project.

References

1. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007)
2. Brants, T., Franz, A.: Web 1T 5-gram corpus version 1.1. Linguistic Data Consortium, Philadelphia (2006)
3. Buitelaar, P., Cimiano, P. (eds.): *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. Frontiers in Artificial Intelligence and Applications, vol. 167. IOS Press, Amsterdam (2008)
4. Buitelaar, P., Cimiano, P., Magnini, B. (eds.): *Ontology Learning from Text: Methods, Evaluation and Applications*. Frontiers in Artificial Intelligence and Applications, vol. 123. IOS Press, Amsterdam (2005)
5. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

6. Curran, J.R., Moens, M.: Improvements in automatic thesaurus extraction. In: Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition, Philadelphia, Pennsylvania, USA, July 2002, pp. 59–66. Association for Computational Linguistics (2002)
7. Dagan, I.: Contextual Word Similarity, ch. 19, pp. 459–476. Marcel Dekker Inc., New York (2000)
8. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6), 391–407 (1990)
9. Fellbaum, C.: WordNet. An Electronic Lexical Database. MIT Press, Cambridge (1998)
10. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1606–1611 (2007)
11. Giuliano, C., Gliozzo, A., Strapparava, C.: Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In: Proceedings of the Fourth International Workshop on Semantic Evaluations, SemEval 2007, Prague, Czech Republic, June 2007, pp. 145–148 (2007)
12. Giuliano, C., Gliozzo, A.M., Strapparava, C.: Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In: Fourth International Workshop on Semantic Evaluations, SemEval 2007. ACL (2007)
13. Gliozzo, A.: Semantic Domains in Computational Linguistics. PhD thesis, University of Trento (2005)
14. Gliozzo, A., Pennacchiotti, M., Pantel, P.: The domain restriction hypothesis: Relating term similarity and semantic consistency. In: Proceedings of NAACL-HLT (2006)
15. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France (July 1992)
16. Ito, M., Nakayama, K., Hara, T., Nishio, S.: Association thesaurus construction methods based on link co-occurrence analysis for wikipedia. In: CIKM 2008: Proceeding of the 17th ACM conference on Information and knowledge management, pp. 817–826. ACM, New York (2008)
17. McCarthy, D., Navigli, R.: Semeval-2007 task 10: English lexical substitution task. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, June 2007, pp. 48–53. Association for Computational Linguistics (2007)
18. Milne, D., Witten, I.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence, WIKIAI 2008 (2008)
19. Nakayama, K., Hara, T., Nishio, S.: A thesaurus construction method from large scaleweb dictionaries. In: 21st International Conference on Advanced Networking and Applications (AINA 2007), pp. 932–939 (2007)
20. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: AAAI. AAAI Press, Menlo Park (2006)
21. Suchanek, F., Kasneci, G., Weikum, G.: Yago - a large ontology from wikipedia and wordnet. *Elsevier Journal of Web Semantics* 6(3), 203–217 (2008)
22. Yang, D., Powers, D.M.: Automatic thesaurus construction. In: Dobbie, G., Mans, B. (eds.) Thirty-First Australasian Computer Science Conference (ACSC 2008), CRPIT, Wollongong, NSW, Australia, vol. 74, pp. 147–156. ACS (2008)