

Automatic Association of Web Directories to Word Senses

Celina Santamaría
Julio Gonzalo
Felisa Verdejo *
UNED, Madrid

We describe an algorithm that combines lexical information (from WordNet 1.7) with Web directories (from the Open Directory Project) to associate word senses with such directories. Such associations can be used as rich characterizations to acquire sense-tagged corpora automatically, cluster topically-related senses and detect sense specializations.

The algorithm is evaluated for the 29 nouns (147 senses) used in the Senseval 2 competition, obtaining 148 (word sense, Web directory) associations covering 88% of the domain-specific word senses in the test data with 86% accuracy.

The richness of Web directories as sense characterizations is evaluated in a supervised Word Sense Disambiguation task using the Senseval 2 test suite. The results indicate that, when the directory/word sense association is correct, the samples automatically acquired from the Web directories are nearly as valid for training as the original Senseval 2 training instances.

The results support our hypothesis that Web directories are a rich source of lexical information: cleaner, more reliable and more structured than the full Web as a corpus.

Introduction

Combining the size and diversity of the textual material in the World Wide Web with the power and efficiency of current search engines is an attractive possibility to acquire lexical information and corpora. A widespread example is spell-checking: many Web users routinely use search engines to assess which is the “correct” (i.e. with more hits in the Web) spelling of words. Among NLP researchers, Web search engines have already been used as a point of departure for:

Extraction of parallel corpora In (Resnik, 1999; Nie, Simard, and Foster, 2001; Ma and Liberman, 1999; Resnik and Smith, 2002), the Web is harvested in search of pages which are available in two languages, with the aim of building parallel corpora for any couple of target languages. This is a very promising technique, as many Machine Translation (MT) and Cross-Language Information Retrieval (CLIR) strategies rely on the existence of parallel corpora, which are still a scarce resource. Such Web-mined parallel corpora have proved to be useful, for instance, in the context of the CLEF (Cross-Language Evaluation Forum) CLIR competition, where many participants use such parallel corpora (provided by the University of Montreal) to improve the performance of their systems (Peters et al., 2002).

* ETS Ingeniería Informática de la UNED, c/ Juan del Rosal, 16, Ciudad Universitaria, 28040 Madrid, Spain

Automatic acquisition of sense-tagged corpora. The description of a word sense can be used to build rich queries in a way that the occurrences of the word in the documents retrieved are, with some probability, associated to the desired sense. If the probability is high enough, it is then possible to acquire sense-tagged corpora in a fully automatic fashion. Again, this is an exciting possibility that would solve the current bottleneck of supervised Word Sense Disambiguation (WSD) methods (namely, that sense-tagged corpora are very costly to acquire). One example of such kind of techniques is (Mihalcea and Moldovan, 1999a), where a precision of 91% is reported over a set of 20 words with 120 senses. In spite of the high accuracy obtained, such methodology did not perform well in the comparative evaluation reported in (Agirre and Martínez, 2000), perhaps indicating that examples obtained from Web may have topical biases (depending on the word), and that further refining is required. For instance, a technique that behaves well with a small set of words might fail in the common cases where a new sense is predominant on the Web, e.g. “oasis” or “nirvana” as music groups, “tiger” as a golfer, “jaguar” as a car brand, etc.

Extraction of lexical information In (Agirre et al., 2000), search engines and the Web are used to assign Web documents to WordNet concepts. Such sets of documents are then processed to build *topic signatures*, i.e. sets of words with weights that enrich the description of a concept.

In (Grefenstette, 1999), the number of hits in Web search engines is used as a source of evidence to select optimal translations for multi-word expressions. For instance, “*apple juice*” is selected as a better translation than “*apple sap*” for the German *ApfelSaft* because “*apple juice*” hits a thousand times more documents in Altavista.

Finally, in (Joho and Sanderson, 2000) and (Fujii and Ishikawa, 1999), the Web is used as a resource to provide descriptive phrases or definitions for technical terms.

A common problem to all the above applications is how to detect and filter out all the noisy material in the Web, and how to characterize the rest (Kilgarriff, 2001b).

Our starting hypotheses is that Web directories (e.g. Yahoo, Altavista or Google directories, the Open Directory Project -ODP-, etc.), where documents are mostly manually classified in hierarchical topical clusters, are an optimal source to acquire lexical information; its size is not comparable to the full Web, but they are anyway huge sources of semi-structured, semi-filtered information waiting to be mined.

In this paper, we describe an algorithm to assign Web directories (from the Open Directory Project, <http://dmoz.org>) as characterizations for word senses in WordNet 1.7 noun synsets (Miller, 1990). For instance, let us consider the noun *circuit*, which has six senses in WordNet 1.7. These senses are grouped in *synsets* together with their synonym terms, and linked to broader (more general) synsets via hypernymy relations:

6 senses of circuit

Sense 1: {circuit, electrical circuit, electric circuit} => {electrical device}

Sense 2: {tour, circuit} => {journey, journeying}

Sense 3: {circuit} => {path, route, itinerary}

Sense 4: {circuit (judicial division)} => {group, grouping}

Sense 5: {racing circuit, circuit} => {racetrack, racecourse, raceway, track}

Sense 6: {lap, circle, circuit} => {locomotion, travel}

Our algorithm associates *circuit 1* (electric circuit) with ODP directories such as:

```
business/industries/electronics and electrical/contract manufacturers
```

while the *circuit 5* (racing circuit) is tagged with directories such as:

```
sports/motorsports/auto racing/tracks
sports/equestrian/racing/tracks
sports/motorsports/auto racing/formula one
```

Every ODP directory has an associated URL, which contains a description of the directory and a number of websites that have been manually listed as pertaining to the directory topic, accompanied by brief descriptions of every site. This information is completed with a list of subdirectories, each of them containing more Web sites and sub-directories. Finally, some directories have also pointers to the same category in other languages. For instance, the Web page for the directory `sports/motorsports/auto racing/tracks` can be seen in Figure 1. This directory contains links and descriptions for 846 Web sites organized in 12 subdirectories, a link to a related directory (`sports/motorsports/karting/tracks`) and a link to the same category in French.

The association of word senses with Web directories is related to the assignment of domain labels to WordNet synsets as described in (Magnini and Cavaglia, 2000), where WordNet is (manually) enriched with domain categories from the Dewey Decimal Classification (DDC). Some clear differences are that directories from the ODP are assigned automatically, are richer, deeper, and, more importantly, come with a large amount of associated information directly retrievable from the Web. DDC categories, on the other hand, are a stable domain characterization compared to Web directories.

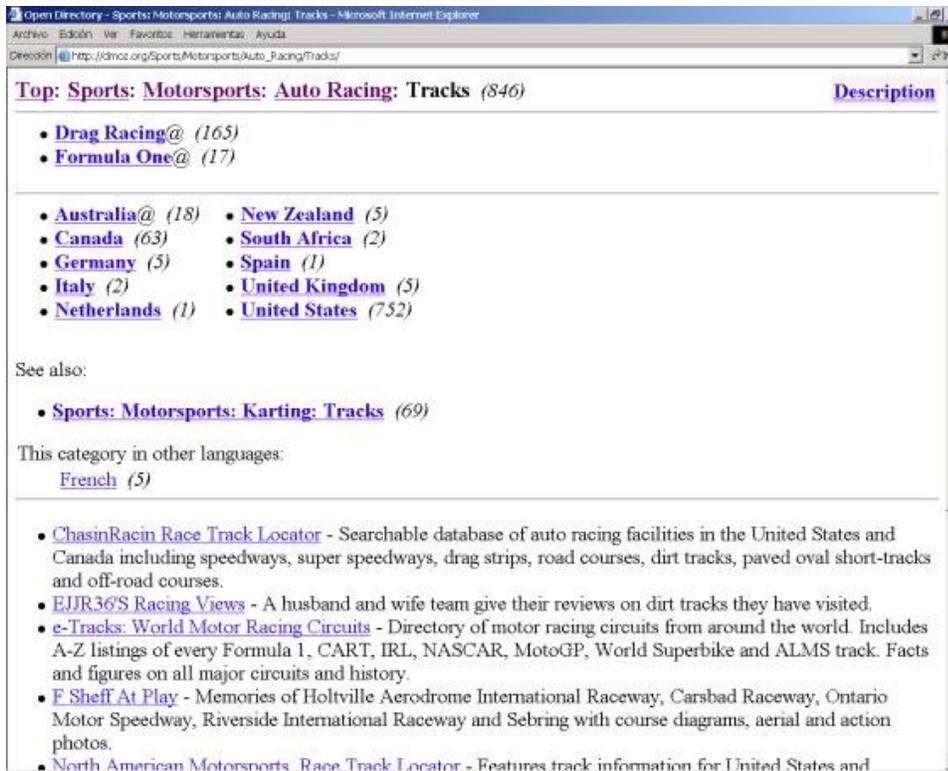
As WordNet and ODP are both hierarchical structures, connecting them is also related to research in mapping thesauri for Digital Libraries, ontologies and data structures in compatible databases. A salient feature of our task is, however, that we do not intend to map both structures, as they are of a quite different nature (lexicalized English concepts vs. topics on the Web). Our goal is rather to associate individual items in a many-to-many fashion. A word sense may be characterized with several Web directories, and a Web directory may be suitable for many word senses.

The most direct applications of word sense/Web directory associations are:

- Clustering of senses with identical or very similar categories.
- Refinement of senses into specialized variants (e.g. *equestrian circuit*, *formula one circuit* as specializations of *racing circuit* in the example above).
- Extraction of sense-tagged corpora from the Web sites listed under the appropriate directories.

In Section 1 we describe the proposed algorithm. In Section 2, we evaluate the precision and recall of the algorithm for the set of nouns used in the Senseval 2 WSD competition. In Section 3, we make a preliminary experiment using the material from ODP directories as training corpora for a supervised WSD system. In section 4, we present the results of applying the algorithm to most WordNet 1.7 nouns. Finally, in Section 5 we draw some conclusions.

Figure 1
Contents of an ODP Web directory associated to *circuit* 5 (racing circuit).



1 Algorithm

Overall, the system takes a WordNet 1.7 noun as input, generates and submits a set of queries into the Open Directory Project (ODP), filters the information obtained from the search engine and returns a set of ODP directories classified as: a) pseudo domain labels for some word sense, b) noise and c) salient noise (i.e. directories which are not suitable for any sense in WordNet, but could reveal and characterize a new relevant sense of the noun). In case a), the WordNet sense \leftrightarrow ODP directory association also receives a probability score.

A detailed description of the algorithm steps follows.

1.1 Querying ODP structure

For every sense w_i of the noun w , a query q_i is generated, including w as compulsory term, the synonyms and direct hypernyms of w_i as optional terms, and the synonyms of other senses of w as negated (forbidden) terms. These queries are submitted to ODP, and a set of directories is retrieved. For instance, for *circuit* the following queries are

generated and sent to the ODP search engine: ¹

```
q1= [+circuit "electrical circuit" "electric circuit" "electrical device" -tour
-"racing circuit" -lap -circle]
q2= [+circuit tour journey journeying -"electrical circuit" -"electric circuit"
-"electrical device" -"racing circuit" -lap -circle]
q3= [+circuit path route itinerary -"electrical circuit" -"electric circuit"
-"electrical device" -tour -"racing circuit" -lap -circle ]
q4= [+circuit group grouping -"electrical circuit" -"electric circuit"
-"electrical device" -tour -"racing circuit" -lap -circle]
q5= [+circuit "racing circuit" racetrack racecourse raceway track -"electrical circuit"
-"electric circuit" -"electrical device" -tour -lap -circle]
q6= [+circuit lap circle locomotion travel -"electrical circuit" -"electric circuit"
-"electrical device" -tour -"racing circuit" -lap -circle]
```

1.2 Representing retrieved directory descriptions

For every directory d , a list of words $l(d)$ is obtained removing stop words and preserving all content words in the directory path. For instance, one of the directories produced by the *circuit* queries is:

```
 $d$  =business/industries/electronics and electrical/contract manufacturers
which is characterized by the following word list:
 $l(d)$  = [business, industries, electronics, electrical, contract, manufacturers]
```

1.3 Representing WordNet senses

For every sense w_j , a list $l(w_j)$ of words is made with

- all nouns in the hypernym chain of maximal length 6
- all hyponyms
- all meronyms, holonyms and coordinate terms

of w_j in WordNet. $l(w_j)$ is used as description of the sense w_j . For instance, *circuit 1* receives the following description:

```
 $l(circuit_1)$  = [electrical circuit, electric circuit, electrical device, bridge,
bridge circuit, wheatstone bridge, bridged-T, closed circuit, loop, parallel circuit,
shunt circuit, computer circuit, gate, logic gate, AND circuit, AND gate, NAND circuit,
NAND gate, OR circuit, OR gate, X-OR circuit, XOR circuit, XOR gate, integrated circuit,
(..)
instrumentality, instrumentation, artifact, artefact, object, physical object, entity]
```

1.4 sense/directory comparisons

For every sense description $l(w_j)$, a comparison is made with the terms in the directory description $l(d)$. This comparison is based on the hypothesis that the terms in an appropriate directory for a word sense will have some correlation with the sense description via WordNet semantic relations. In other words, our assumption is that the path to the directory in the ODP topical structure will have some degree of overlapping with the hyponymy path to the word sense in the WordNet hierarchical structure.

For this comparison, we simply count the number of co-occurrences between words in $l(w_j)$ and words in $l(d)$. Repeated terms are not discarded, as repetition is correlated with

¹ In ODP queries, compulsory terms are denoted by + and forbidden terms by -

stronger associations. Other, better grounded comparisons, such as the cosine between $l(w_j)$ and $l(d)$, were empirically discarded due to the small size and little overlapping of the average vectors.

1.5 Candidate sense/directory associations

The association vector $v(d, w)$ has as many components as senses for w in WordNet 1.7; the i^{th} component, $v(d, w)_i$ represents the number of matches between the directory $l(d)$ and the sense descriptor $l(w_j)$. For instance, the association vector of

`business/industries/electronics` and `electrical/contract manufacturers`
with *circuit* is

$$v(d, \text{circuit}) = (6, 0, 0, 0, 0, 0)$$

i.e. six coincidences for sense 1 (the “*electric circuit*” sense), which has the associated vector shown in the previous section (that includes five occurrences of *electrical* and one occurrence of *electronic*). The rest of the sense descriptions have no coincidences with the directory description.

$v(d, w)$ is the basis to make candidate assignments of suitable senses for directory d : If one of the components $v(d, w)_j$ is not null, we assign the sense w_j to the directory d . If all components are null, the directory is provisionally classified as noise or new sense. If more than one component is not null, the senses i with maximal $v(d, w)_i$ are all considered as candidates.

These candidate assignments are confirmed or discarded after passing a number of filters and receiving a confidence score $C(d, w_j)$, which are described below.

1.6 Filters

Filters are simple heuristics that contribute to a more accurate classification of the relations predicted by the cooccurrence vector $v(d, w)$. We are currently using two filters: one differentiates nouns and noun modifiers to prevent wrong associations and another detects sense specializations.

1.6.1 Modifiers Frequently, the ODP search engine retrieves directories where the noun to be searched, w , has as a noun modifier role. Such cases usually produce wrong associations. For instance, the directory:

`library/sciences/animals & wildlife/mammals/tamarins/golden lion tamarin`
is erroneously associated to the mammal sense of “lion”, which is here a modifier for “tamarin”.

Modifiers are detected with a set of simple patterns, as the syntactic properties of descriptions in directories are quite simple. In particular, we discard most cases using the structure of the ODP hierarchy, as in this case. The filter analyses the structure of the directory, and detects that the parent category of *golden lion tamarin* is *tamarin*, and therefore assumes that *golden lion tamarin* is a specialization of *tamarin*, and assigns the directory to a suitable sense of *tamarin* (*tamarin 1* in WordNet).

An additional filter (weaker than the previous one) is discarding compounds according to the position (the searched noun precedes another noun), as in:

`personal/kids/arts & entertainment/movies/animals/lion king`

This directory could be associated to *lion 1* because it contains the word *animal*, but the assignment is rejected due to the modifier filter. In general, on such occasions the searched noun plays a modifier role (as adjective or noun); discarding all such cases favors precision against recall. In this case, the label is classified as noise.

1.6.2 Sense specializations (hyponyms) A retrieved directory might be appropriate as a characterization of a sense specialization for some of the word senses being considered; our algorithm tries to detect such cases, creating a hyponym of the sense and characterizing the directory with the hyponym.

The filter identifies a directory as a candidate hyponym if it contains explicitly a **modifier w** pattern (where w is the noun being searched). This filter detects explicit specializations such as *office chair* as a hyponym of *chair 1*, or *family fox channel* as a hyponym of *channel 7*; but fails to identify, for instance, *memorial day* as a hyponym of *holiday*.

If the candidate hyponym, as a compound, is not present in WordNet, then it is incorporated and described with the directory. If it is already present in WordNet, an additional checking of the hyponymy relation is made. For instance, the directory

`business/industries/electronics` and `electrical/components/integrated circuits` is assigned to the WordNet entry *integrated circuit*, because *integrated circuit* is already a hyponym of *circuit* in WordNet.

1.7 Confidence score

Finally, a confidence score $C(d, w_j)$ for every potential association (d, w_j) is calculated using four empirical criteria:

1. Checking whether d was directly retrieved for the query associated to w_j .
2. Checking whether the system associates d to one or more senses of the word w .
3. Checking the number of coincidences between $l(d)$ and $l(w_j)$.
4. Comparing the previous number with the number of coincidences between $l(d)$ and the other sense descriptions $\{l(w)_i, i \neq j\}$.

The confidence score is a linear combination of these factors, weighted according to an empirical estimation of their relevance:

$$C(d, w_j) = \sum_{i=1}^4 \alpha_i C_i(d, w_j)$$

where

$$C_1(d, w_j) = \begin{cases} 1 & \text{if query}(w_j) \text{ retrieves } d \\ 0 & \text{otherwise} \end{cases}$$

$$C_2(d, w_j) = 1 - \frac{k}{n}$$

$$C_3(d, w_j) = \begin{cases} 1, & \text{if } v_j \geq 5 \\ (v_j + 5)/10, & \text{if } 1 < v_j \leq 4 \\ 0.5, & \text{if } v_j = 1 \end{cases}$$

$$C_4(d, w_j) = \frac{v_j - \max_{i \neq j}(v_i)}{\sum_{i=1}^n v_i}$$

being v the association vector $v(d, w)$, n the number of senses, k the number of senses for which v_j is non-null, and α_i coefficients empirically adjusted to $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.1, 0.15, 0.4, 0.35)$. The value of $C(d, w_j)$ ranges between 0 and 1 (all C_i range between 0 and 1, and the sum of the linear coefficients α_i is 1). Note that C_2 cannot reach 1 (but

can get asymptotically close to 1), and note also that C_4 cannot take negative values, because - as (d, w_j) is a candidate association, v_j is maximal in $v(d, w)$ and therefore $v_j - \max_{i \neq j}(v_i)$ ranges between 0 and v_j .

Let us see an example of how this confidence measure works, calculating $C(d, w_j)$ for the directory `d=business/industries/electronics` and `electrical/contract manufacturers` with *circuit 1* (electric circuit):

- C_1 . This directory has been retrieved from the query

```
q1= [+circuit "electrical circuit" "electric circuit" "electrical device" -tour
-"racing circuit" -lap -circle]
```

corresponding to *circuit 1*, which agrees with the association made by the system. Hence $C_1 = 1$.

- C_2 . The association vector $v(d, w) = (6, 0, 0, 0, 0)$ presents only one non null coordinate; Therefore $C_2 = 1 - \frac{1}{6} = 0.83$. Note that, in general, this factor prevents C from reaching the upper bound 1.
- C_3 . As $v_1 = 6$, $C_3 = 1$. This factor increases along with the number of coincidences between the sense and directory characterizations.
- C_4 . As all other components of v are null, the highest value of the components different from sense 1 is also null ($\max_{i \neq j}(v_i) = 0$); therefore, $C_4 = 1$. This factor measures the strength of the association (d, w_1) compared with the other possibilities. It decreases when $v(d, w)$ includes more than one non null coordinate, and their values are similar.
- C . Finally, applying the α_i coefficients, we obtain $C(d, \textit{circuit 1}) = 0.975$

The confidence score can be used to set a threshold for accepting/discarding associations; a higher threshold should produce a lower number of highly precise associations; a lower threshold would produce more associations with less accuracy. For the evaluation below, we have retained all directories, regardless of their confidence score, in order to assess how well this empirical measure correlates with correct and useful assignments.

An example of the results produced by the algorithm can be seen in Figure 2. The system assigns directories to senses 1, 2 and 5 of *circuit* (6, 2 and 3 directories, respectively). Some of them are shown in the figure, together with a sense specialization, *integrated circuit* for sense 1 (electrical circuit). Senses 3, 4 and 6, which did not receive any directory association, do not appear to have domain specificity, but are rather general terms.

2 Evaluation

We have analyzed the results of the algorithm for the set of nouns in the Senseval 2 WSD English lexical sample testbed (Kilgarriff, 2001a). The Senseval campaigns (Edmonds and Cotton, 2001; Kilgarriff and Palmer, 2000) are devoted to the comparative evaluation of Word Sense Disambiguation systems in many languages. In the Senseval 2 *lexical sample* task, a large number of instances (occurrences in context extracted from corpora) for a fixed set of words had to be tagged with the appropriate sense by the participating WSD systems. For English, the sense inventory was a pre-release of WordNet 1.7, and two sets of manually tagged instances were made available: a first set was intended for training supervised systems, and a second set for evaluation of all systems attempting the task. Altogether, the Senseval 2 lexical sample testbed is one of the most widely used resources to study and compare Word Sense Disambiguation approaches.

Figure 2Results of the association algorithm for *circuit*

circuit 1 (electrical circuit)	
<i>ODP directories</i>	<i>C</i>
business/industries/electronics and electrical/contract manufacturers	0.98
manufacturers/printed circuit boards/fabrication	0.88
computers/cad/electronic design automation	0.78
...	
<i>sense specializations (hyponyms)</i>	
business/industries/electronics and electrical/components/integrated circuits	0.98

circuit 2 (tour, journey around a particular area)	
<i>ODP directories:</i>	
sports/cycling/travel/travelogues/europe/france	0.58
regional/asia/nepal/travel and tourism/travel guides	0.66

circuit 5 (racing circuit)	
<i>ODP directories:</i>	
sports/motorsports/auto racing/stock cars/drivers and teams	0.78
sports/motorsports/auto racing/tracks	0.82
sports/motorsports/auto racing/driving schools	0.78

For our evaluation, we have considered the fraction of the Senseval 2 testbed that deals with English nouns: 29 polysemous nouns with a total of 147 word senses. We have applied the algorithm to this set of nouns, and examined the results in terms of coverage and quality of the sense/directory associations. Coverage measures how many senses can be characterized with directories, assuming that every domain-specific sense should receive at least one directory. Quality is measured in terms of precision (are the assignments correct?), relevance (are the assignments useful?), and confidence (does the confidence score correlate well with precision and relevance of the associations?).

2.1 Coverage

Table 1 shows the 148 directories retrieved by our algorithm, which makes an average of 1.0 directories per sense. The directories, however, are not evenly distributed among senses, covering only 43 different senses with unique directories, and 28 specialized (hyponym) senses. In addition, 9 senses are identified as part of potential clusters (i.e. having non-unique directories).

In order to measure the real coverage of the system, we have to estimate how many word senses in the Senseval 2 sample are susceptible to receive a domain label. For instance, “*sense*” in “*common sense*” is not associated to any particular topic or domain, while “*sense*” as “*word sense*” can be associated to linguistics or language-related topics.

The decision of whether a word sense might receive a domain label or not is not always a simple, binary decision. Hence we have manually tagged all word senses with two criteria (each performed by a different human annotator): a strict one (only word senses which can clearly receive a domain label are marked as positive), and a loose

Table 1

Coverage of nouns in the Senseval 2 testbed.

Senseval 2 nouns	# senses	# directories	# labeled senses	# hyponyms
art	4	6	1	1
authority	7	4	2	1
bar	13	3	3	0
bum	4	0	0	0
chair	4	4	1	0
channel	7	5	1	1
child	4	12	2	0
church	3	24	2	4
circuit	6	11	3	1
day	10	15	1	14
detention	2	1	1	0
dyke	2	1	1	0
facility	5	10	3	0
fatigue	4	0	0	0
feeling	6	2	1	0
grip	7	3	2	0
hearth	3	5	2	0
holiday	2	2	2	0
lady	3	0	0	0
material	5	9	2	3
mouth	8	0	0	0
nation	4	4	1	1
nature	5	0	0	0
post	8	14	5	0
restraint	6	4	3	0
sense	5	0	0	0
spade	3	3	1	1
stress	5	5	2	1
yew	2	1	1	0
Total	147	148	43	28

one (only word senses which are completely generic are marked as negative). The strict judgment gave 59 domain-specific senses in the sample; the loose judgment gave 71.

With these manual judgments, the coverage of the algorithm is between 73% (loose judgment) and 88% (strict judgment). This coverage can be increased by:

- Propagating a directory / word sense association to all members of the WordNet synset to which the word sense belongs.
- Propagating directories via hyponymy chains as in (Magnini and Cavaglia, 2000).

2.2 Quality

We have used three criteria to evaluate the directory/sense associations produced:

Precision Is the directory *correct* (suitable) for the word sense or not?

Relevance Is the directory *useful* to characterize the word sense?

Confidence How well is the confidence value $C(d, w_j)$ correlated with precision and relevance of the associations?

2.2.1 Precision An assignment (d, w_j) is considered correct (d is suitable for w_j) unless:

1. d adjusts better to some other sense w_i . For instance, the association of:

`regional/north america/united states/government/agencies/independent/federal
labor relations authority`

as a hyponym of

authority 4: assurance, self-assurance, confidence, self-confidence, authority, sureness

is considered an error, as the directory would be better suited for a hyponym of sense 5 (authority as *administrative unit*).

2. The terms in $l(d)$ are contradictory with the definition of the word sense, or are better suited for a sense that is not listed in the dictionary. This is the case of:

`arts/music/bands and artists/offspring`

which is erroneously assigned to *child 2: human offspring of any age*.

The results of this manual evaluation can be seen in Table 2. The overall precision is 86%.

Regarding potential topical clusters (directories associated to more than one sense of the same word), these are considered correct if a) the associated directory is correct for all the senses in the cluster, and b) the occurrences of the word in the Web page associated to the directory can be loosely assigned to any of the cluster senses. 12 out of the 13 clusters extracted are correct according to this criterion.

2.2.2 Confidence measures Table 3 shows the distribution of directories according to the confidence measure. 84% of the directories have a confidence C over 0.7, and 41% over 0.8. This skewed distribution is consistent with the algorithm filters, designed to favor precision rather than recall.

Table 4 shows the distribution of errors in levels of confidence. The percentage of errors in directories with a confidence below .6 is 25%. This error percentage decreases with increasing levels of confidence, down to 5% for associations with C over .8. This table indicates that the confidence value, which is assigned heuristically, is indeed correlated with precision.

Table 2
Precision over Senseval 2 nouns.

Directories associated to WordNet senses	# directories	# correct	# errors
Unique sense	148	127	21
potential clustering	13	12	1
Total	161	139 (86%)	22 (14%)

Table 3

Confidence distribution.

Confidence	$C \leq 0.7$	$0.7 < C \leq 0.8$	$0.8 < C$
# directories	24	63	61

Table 4

Correlation between confidence and correctness

Confidence	# directories	% errors
$C \leq 0.7$	24	25%
$0.7 < C \leq 0.8$	63	19%
$C > 0.8$	61	5%
Total	148	14%

2.2.3 Relevance Besides correctness of the associations, we want to measure the usefulness of the directories: how well they can be used to characterize the associated word senses? How much information do they provide about the word sense?

We have done a manual, qualitative classification of the directories extracted as *irrelevant*, *mildly relevant* or *very relevant*. An irrelevant directory is compatible with the word sense, but does not provide any useful characterization; a mildly relevant directory illustrates the word sense, but not centrally or in some particular aspect or domain. A very relevant directory provides a rich characterization per se, and can be considered as a domain label for the word sense.

An example of a very relevant directory is:

`business/industries/electronics and electrical/components/integrated circuit`

associated as hyponym of *circuit 1 (electrical circuit)* with a confidence of 98%.

An example of mildly relevant association is

`regional/north america/united states/texas/./society and culture/religion`

associated to *church 1 (Christian church)* with a 73% confidence. Obviously, Texas is not correlated with church, but the directory contains a lot of material (for instance the Web page of the “Northcrest Community Church” and many others) that might be used, for instance, to acquire topical signatures for the concept. Hence the “mildly relevant” judgment.

Finally, an irrelevant association is:

`regional/north america/united states/new york/localities/utica`

associated to *art 1 (fine art)* with a confidence of 66% (the directory contains a section of Arts at Utica, which would be considered mildly relevant if pointed explicitly by the label).

For the purposes of measuring relevance, all the directories that were judged as *incorrect* are counted as *irrelevant*.

The overall relevance figures, and the correlation of relevance with the confidence value, can be seen in Table 5. 67% of the directories are highly relevant to characterize word senses, which is an encouraging result. Also, the set of irrelevant directories (15%)

is almost identical to the set of erroneous directories (with just one addition), indicating that (almost) all directories that are correct can be used to characterize word senses to some extent.

Table 5

Relevance of the directories in the test set.

Relevance	Irrelevant	Mildly relevant	Highly relevant
$C \leq 0.7$	7	4	13
$0.7 < C \leq 0.8$	13	12	38
$0.8 < C$	3	9	49
Total	23 (15%)	25 (17%)	100 (67%)

3 Example application: automatic acquisition of sense-tagged corpora

Each ODP directory contains links to related subdirectories, and to a large number of Web sites that have been manually classified there. Every link includes the name of the Web site and a short description. For instance, under

`business/industries/electronics and electrical/components/integrated circuit`

we find over 30 descriptions such as ‘Multilink Technology corporation: Manufacture of integrated circuits, modules, and boards for use in both data and telecommunications’.

In order to perform a first experiment on extraction of sense-tagged corpora, we have used only such descriptions (without exploring the associated Web sites) to build a sense-tagged corpus for Senseval 2 nouns.

Notice that we are not using the contents of the Web sites that belong to a directory, but only the manually added description of Websites in the directory. Using the Web sites themselves is also an attractive possibility, that would produce a much larger corpus at the expense of lower precision.

The extraction is straightforward: when a word sense w_i has an associated directory d , we scan the site descriptions in the ODP page that corresponds to the directory d and extract all contexts in which w occurs, assuming that in all of them w is used in the sense i .

Some examples of the training material for *circuit* can be seen in Figure 3. On average, these examples are shorter than Senseval 2 training instances.

The goal is to compare the performance of a supervised Word Sense Disambiguation system using Senseval 2 training data (hand made for the competition) versus using the sense-tagged corpus from ODP (automatically extracted). We have chosen the *Duluth* system (Pedersen, 2001) to perform the comparison. The Duluth system is a freely available supervised WSD system that participated in the Senseval 2 competition. As we are not concerned with absolute performance, we simply adopted the first of the many available versions of the system (*Duluth 1*).

An obstacle to perform such comparative evaluation is that, as expected, our algorithm assigns ODP directories only to a fraction of all word senses, partly because not every sense is domain-specific, partly due to lack of coverage. In order to circumvent this problem, we have considered only the subset of 10 Senseval 2 nouns for which our system tags at least two senses: *bar*, *child*, *circuit*, *facility*, *grip*, *holiday*, *material*, *post*, *restraint*, *stress*. We have then projected the Senseval 2 training corpus, and the test material, to the annotations for the word senses already in our ODP-based material. Hence

Figure 3Examples of training material for *circuit***circuit 1 (electrical circuit)**

Electromechanical products for brand name firms; offers printed circuit boards (..)

Offers surface mount, thru-hole, and flex circuit assembly, in circuit and functional (..)

circuit 2 (tour, journey around a particular area)

The Tour du Mont-Blanc is a circuit of 322km based in the northern French Alps.

A virtual tour of the circuit by Raimon Bach.

circuit 5 (racing circuit)

The Circuit is a smooth 536 yards of racing for Hot Rod and Stock Car's at the East of (..)

(..) History of the circuit and its banked track and news of Formula 1 (..)

we will evaluate the quality of the training material obtained from Web directories, not the coverage of the approach.

Table 6 shows the training material obtained for that subset of Senseval 2 nouns. A total of 66 directories are used as a source of training instances, from which 17% of them are incorrect and will presumably incorporate noise into the training.

Table 6

Training material obtained for the WSD experiment

word senses	# directories per sense	# incorrect directories	# training instances
bar 1,10	1,1	0,0	1,1
child 1,2	3,9	0,0	3,80
circuit 1,2,5	6,2,3	0,0,0	229,2,5
facility 1,4	4,5	0,0	4,18
grip 2,7	2,1	0,1	17,6
holiday 1,2	1,1	0,1	5,17
material 1,4	6,3	2,1	63,10
post 2,3,4,7,8	1,5,1,4,3	1,1,1,0,3	2,7,1,9,3
restraint 1,4,6	2,1,1	0,0,0	2,2,2
stress 1,2	1,4	0,0	8,50
Total	66	11	547

Table 7 compares the training material for the word senses in this sample, and the results of the supervised WSD algorithm with the Senseval and the ODP training instances.

We have measured the performance of the system in terms of Senseval *recall*: the number of correctly disambiguated instances over the total number of test instances. Overall, using the Senseval training set gives .73 recall, and training with the automati-

Table 7
Results of supervised WSD

word senses	# instances Senseval training	# instances ODP training	# test instances	Recall senseval training	Recall ODP training
bar 1,10	127,11	1,1	62,6	.91	.50
child 1,2	39,78	3,80	35,27	.57	.44
circuit 1,2,5	67,6,7	229,2,5	23,2,8	.70	.70
facility 1,4	26,61	4,18	15,28	.79	.67
grip 2,7	6,1	17,6	4,0	1	1
holiday 1,2	4,57	5,17	26,2	.96	.96
material 1,4	65,7	63,10	30,9	.79	.79
post 2,3,4,7,8	1,64,20,11,7	2,7,1,9,3	2,25,13,12,4	.45	.25
restraint 1,4,6	17,32,11	2,2,2	8,14,4	.65	.50
stress 1,2	3,45	8,50	1,19	.95	.95
Total	773	547	379	.73	.58

cally extracted ODP instances gives .58 (21% worse). A decrease of 21% is significant but nevertheless encouraging, because the Senseval training set is the gold standard for the Senseval test set: it is larger than the ODP set (773 versus 547 instances in this subset), well balanced, built with redundant manual annotations, and part of the same corpus as the test set.

The most similar experiment in the literature is (Agirre and Martínez, 2000), where the sense-tagged instances obtained using a high-performance Web mining algorithm (Mihalcea and Moldovan, 1999a) performed hardly better than a random baseline as WSD training instances. A difference between the two experiments is that Agirre et al. do not limit their experiments to the fraction of the test set for which they have automatically extracted training samples, hence a direct comparison of the results is not possible.

A detailed examination of the results indicates that the difference in performance is related to the smaller number of training instances rather than to the quality of individual instances:

- In all four cases where ODP provides a comparable -or larger- number of training instances (*circuit*, *grip*, *material*, *stress*), ODP training equals hand-tagged training. In one more case (*holiday*), the number of ODP instances is lower but still the recall is the same. For the other five words, the number of ODP instances is substantially lower and the recall is worse.
- Remarkably, incorrect directories only harm recall substantially for *post*, which accumulates six erroneous associations (out of 11 errors). The other five errors (in *material 1,4*, *holiday 2*, *grip 7*) do not affect the final recall for these words. There are two possible reasons for this behavior:
 - Erroneous directories tend to be less productive in terms of training instances. Indeed, this fact could be incorporated as an additional filter for candidate directories. This is the case, for instance, of *material 1*, for which correct directories provide much more training material than the incorrect one.

- Erroneous directories are more frequent with rare (less frequent) word senses. This is correlated with a lower number of test instances (hence the influence on average recall is lower) and also of training instances (and then the reference, hand-tagged material does not provide good training data either). This is the case of *grip 7* or *holiday 2*, which have 0 and 2 test instances respectively.

Overall, our results suggest that directory-based instances, in spite of being shorter and automatically extracted, are not substantially worse for supervised WSD than the hand-tagged material provided by the Senseval organization. The limitation of the approach is currently the low coverage of word senses and the amount of training samples. Two strategies may help overcoming such limitations: first, propagating directories via synonymy (attaching directories to synsets rather than word senses) and semantic relationships (propagating directories via hyponymy relations); second, retrieving instances not only from the ODP page describing the directory contents, but from the Web pages listed in the directory.

The only fundamental limitation of our approach for the automatic extraction of annotated examples is the fact that directories are closely related to topics and domains and, therefore, word senses that do not pertain to any domain cannot receive directories and training instances from them. Still the approach can be very useful for Language Engineering applications when only domain disambiguation (versus sense disambiguation) is required, such as Information Retrieval (Gonzalo et al., 1998), Content-based User Modelling (Magnini and Strapparava, 2000), etc.

4 Massive processing of WordNet nouns

Table 8

Massive association of ODP directories to WordNet 1.7 nouns

		with propagation
Candidate nouns	51168	
candidate senses	73612	
Associated directories	29291	
Characterized nouns	13375	24558
Characterized senses	14483	27383
Hyponyms	1800	

We have applied the association algorithm to all non-compound nouns in WordNet without non-alphabetic characters (i.e. “sea lion” and “10” are not included in the bulk processing). The results can be seen in Table 8. Overall, the system associates at least one directory to 13375 nouns (28% of the candidate set).

The most direct way of propagating directories in the WordNet structure is extending sense/directory associations to synset/directory relations (i.e., if a word sense receives a directory, then all word senses in the same synset receive the same directory). For instance, *cable 2* (transmission line) receives the following directories:

```
business/industries/electronics and electrical
business/industries/electronics and electrical/hardware/connectors and terminals
```

`business/industries/electronics and electrical/contract manufacturers`

As *cable 2* is part of the synset $\{cable\ 2, line\ 9, transmission\ line\ 1\}$, *line 9* and *transmission line 1* inherit the three directories.

With this (quite conservative) strategy, the number of characterized nouns and word senses almost doubles: 24558 nouns and 27383 senses, covering 34% of the candidate nouns plus 7027 multiword terms that were not in the candidate set.

The results of this massive processing, together with the results for the Senseval 2 test (including training material) are available at <http://nlp.uned.es/ODP> for public inspection.

5 Conclusions

Our algorithm is able to associate ODP directories to WordNet senses with 86% accuracy over the Senseval 2 test, and with coverage between 73% and 88% of the domain specific senses. Such associations can be used as rich characterizations for word senses: as a source of information to cluster senses according to their topical relatedness, to extract topic signatures, to acquire sense-tagged corpora, etc. The only intrinsic limitation of the approach is that Web directories are not appropriate to characterize general word senses (versus domain-specific ones). If such characterization is necessary for a particular Natural Language application, the method should be complemented by other means of acquiring lexical information.

In the supervised WSD experiment we have carried out, the results suggest that the characterization of word senses with Web directories provide cleaner data, without further sophisticated filtering, than a direct use of the full Web. Indeed the WSD results using training material from ODP directories gives better results than could be expected from previous cross-validations of training and test WSD materials.

Our ongoing work is extending the algorithm -which works independently for every input word- to combine and propagate sense/directory associations over the entire WordNet. The initial coverage of WordNet nouns is 34%, but we hope to improve this figure taking advantage of the WordNet structure.

Perhaps the main conclusion of our work is that Web directories are a much more structured and reliable corpus than the whole Web. In spite of being manually supervised, Web directories offer immense amounts of structured corpora that deserve our attention as sources of linguistic information. In particular, listing word sense/ODP directory associations has the additional advantage, compared to other Web mining approaches, of providing a wealth of lexical information in a very condensed manner.

Acknowledgments

This work has been partially supported by the Spanish Government, project *Hermes* (TIC2000-0335-C03-01).

References

- Agirre, E., O. Ansa, E. Hovy, and D. Martínez. 2000. Enriching very large ontologies using the WWW. In *Proceedings of the Ontology Learning Workshop*.
- Agirre, E. and D. Martínez. 2000. Exploring automatic Word Sense Disambiguation with decision lists and the Web. In *Proceedings of the COLING Workshop on Semantic Annotation And Intelligent Content*.
- Chen, J. and J. Nie. 2000. Web parallel text mining for Chinese English Cross-Language Information Retrieval. In *Proc. of the International Conference on Chinese Language Computing*.

- Edmonds, P. and S. Cotton. 2001. Senseval-2: Overview. In *Proceedings of Senseval 2*. Association for Computational Linguistics.
- Fujii, A. and T. Ishikawa. 1999. Utilizing the World Wide Web as an encyclopedia: extracting term descriptions from semi-structured texts. In *Proceedings of ACL-99*. Association for Computational Linguistics.
- Gonzalo, J., F. Verdejo, I. Chugur, and J. Cigarrán. 1998. Indexing with Wordnet synsets can improve text retrieval. In *COLING/ACL'98 Workshop on Usage of WordNet in Natural Language Processing Systems*. Association for Computational Linguistics.
- Grefenstette, G. 1999. The WWW as a resource for Example-Based MT tasks. In *Proceedings of ASLIB-99*.
- Joho, H. and M. Sanderson. 2000. Retrieving descriptive phrases from large amounts of free text. In *Proceedings of the 9th ACM CIKM Conference*.
- Kilgariff, A. 2001a. English lexical sample task description. In *Proceedings of Senseval 2*. Association for Computational Linguistics.
- Kilgariff, A. 2001b. Web as Corpus. In *Proceedings of Corpus Linguistics 2001*.
- Kilgariff, A. and M. Palmer. 2000. Introduction to the Special Issue on Senseval. *Computers and the Humanities*, 34(1-2).
- Ma, X. and M. Liberman. 1999. Bits: a method for bilingual text search over the web. In *Proceedings of the Machine Translation Summit VII*.
- Magnini, B. and G. Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*.
- Magnini, B. and C. Strapparava. 2000. Experiments in word domain disambiguation for parallel texts. In *ACL-2000 Workshop on Word Sense and Multilinguality*. Association for Computational Linguistics.
- Mihalcea, R. and D. Moldovan. 1999a. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI '99*.
- Mihalcea, R. and D. Moldovan. 1999b. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the ACL*. Association for Computational Linguistics.
- Miller, G. 1990. Special issue, Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4).
- Nie, Jian-Yun, Michel Simard, and George Foster. 2001. Multilingual Information Retrieval based on parallel texts from the Web. In *Cross-Language Information Retrieval and Evaluation*. Springer Verlag LNCS Series.
- Pedersen, T. 2001. Machine Learning with lexical features: The Duluth approach to Senseval-2. In *Proceedings of Senseval-2*. Association for Computational Linguistics.
- Peters, C., M. Braschler, J. Gonzalo, and M. Kluck, editors. 2002. *Evaluation of Cross-Language Information Retrieval Systems.*, Springer-Verlag LNCS 2406.
- Resnik, P. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.
- Resnik, P. and N. Smith. 2002. The web as a parallel corpus. Technical Report UMIACS-TR-2002, University of Maryland Technical Report.