Integrating Subject Field Codes into WordNet

Bernardo Magnini and Gabriela Cavaglià

ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica
Loc. Panté di Povo, I-38050 Trento, Italy
magnini@itc.it

Abstract

In this paper, we present a lexical resource where WordNet synsets are annotated with Subject Field Codes. We discuss both the methodological issues we dealt with and the annotation techniques used. A quantitative analysis of the resource coverage, as well as a qualitative evaluation of the proposed annotations, are reported.

1. Introduction

Subject Field Codes (SFC) are sets of relevant words for a specific domain. The best appossimation of SFCs are the field labels used in dictionaries (e.g. MEDICINE, ARCHITECTURE), even if their use is restricted to word usages belonging to specific terminological domains. In WordNet, too, SFCs seem to be used occasionally and without a consistent design.

SFCs are considered a crucial information for a general purpose lexical resource (e.g. the "domain ontology" in the Eurowordnet model) and for many NLP tasks such as in Information Retrieval (IR), Word Sense Disambiguation (WSD) and Text Classification. In IR, SFCs are used to expand a query and "the target" (usually keywords) to include more words in the intersection in order to improve both recall and precision ((Liddy and Paik, 1993) and (Schutze, 1998)). WSD tasks usually compare the possible semantic fields of the ambiguous word with the context of the word in order to identify the most correct senses ((Gythrie et al., 1991), (Yarowsky, 1993), (Wilks and Stevenson, 1998) and (Gonzalo et al., 1998)). Text Classification processes determine the subject area of a text searching for keywords. Because texts may use synonyms or refer to concepts, SFCs may increase the effectiveness of the research by keywords (Walker and Amsler, 1986).

Some NLP works use SFCs derived from pre-existent lexical resources, such as WordNet semantic files (Agirre et al., to appear), the WordNet hierarchy (Hearst and Schutze, 1996) and LDOCE field labels ((Walker and Amsler, 1986), (Gythrie et al., 1991), (Liddy and Paik, 1993) (Wilks and Stevenson, 1998)). But because they are not consistent and complete, in the last years SFCs repertories have been automatically developed. In (Veronis and Ide, 1990), SFCs are automatically derived from MRD definitions using neural networks; in (Dyvik, 1998) SFCs are derived from a multilingual corpus. Some other works such as (Schutze, 1992), (Hearst and Schutze, 1996), (Leacock et al., 1996) and (Schutze, 1998) extract SFCs from corpora using cooccurrence statistical methods.

In this paper, we present a lexical resource where WordNet synsets are annotated with SFCs by a semiautomatic procedure which exploits WordNet structure. The decision to annotate WordNet (Miller et al., 1990) synsets with SFCs is motivated in several respects: (i) SFCs provide cross-categorial information, which is almost completely absent in WordNet; (ii) synsets are the appropriate semantic level for SFC annotation, abstracting from the word level; (iii) as we consider SFCs to be basically language independent, we envisage an important role for them in multilingual wordnet-like resources, such EuroWordNet and MultiWordNet (Artale et al., 1998).

The paper is organized as follow: the SFCs organization is described in 2. while in 3. the procedure for the annotation of the synsets with the appropriate SFCs is described. In 4. and 5. we point out some results and the evaluation process we have used to check the SFCs we have produced.

2. Subject Field Codes organization

Information brought by SFCs is complementary to what is already in WordNet. First of all a SFC may include synsets of different syntactic categories: for instance MEDICINE[1] groups together senses from Nouns, such as `doctor#1` and `hospital#1`, and from Verbs such as `operate#7`.

Second, a SFC may also contain senses from different WordNet sub-hierarchies (i.e. deriving from different "unique beginners" or from different "lexicographer files"). For example, the SPORT SFC contains senses such as `athlete#1`, deriving from `life_form#1`, `game_equipment#1`, from `physical_object#1` `sport#1` from `act#2`, and `playing_field#1`, from `location#1`.

We started deriving a list of about 250 subject field codes from a number of paper and machine readable dictionaries. Then the list has been enriched on the base of the Dewey Decimal Classification (Diekema, 1998), and then structured along two dimensions: inclusion, resulting in a SFC hierarchy, and semantic proximity, resulting in a number of SFC families.

2.1. SFC Hierarchy

As for the hierarchical organization, each level is made up of codes of the same degree of specificity: for example the second level includes SFCs such as BOTANY,

Gabriela Cavaglià is currently at ITRI, Brighton, UK - email: Gabriela.Cavaglia@itri.brighton.ac.uk

[1] Throughout the paper subject field codes are indicated with this TYPEFACE while word senses are reported with this `typeface#1`, with their corresponding numbering in WordNet 1.6.

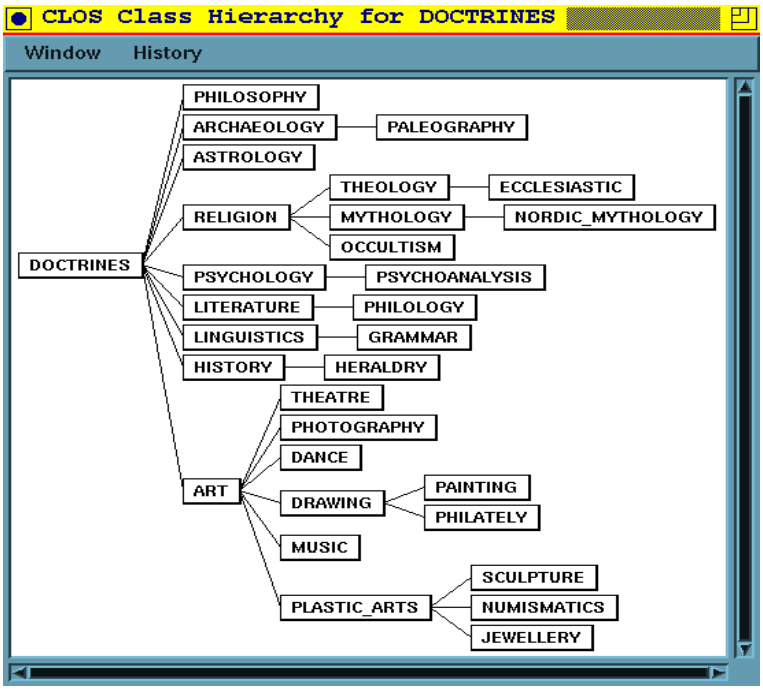

Figure 1: SFC hierarchy fragment.

LINGUISTICS, HISTORY, SPORT and RELIGION, while at the third level we can find specializations such as AMERICAN_HISTORY, GRAMMAR, PHONETICS and TENNIS. Figure 1 shows a fragment of the SFC hierarchy.

2.2. SFC Families

In addition to the hierarchical structure, SFCs are organized into *families*. A family is a group of semantically close SFCs among which there is no inclusion relation. Whereas we consider the hierarchy as a fixed organization, families can be flexibly rearranged, allowing the creation of new interdisciplinary and application dependent codes. As an example, a family such as {SPORT MEDICINE ANATOMY} imposes a particular user's point of view over the SFC organization.

2.3. Factotum and Generics

There are a number of WordNet synsets that do not belong to a specific SFC, but rather they can appear in almost all of them. For this reason, a FACTOTUM SFC has been created which basically includes two types of synsets:

- *Generic* synsets, which are hard to classify in a particular SFC, such as:

  `man#1` an adult male person (as opposed to a woman)
  `man#3` the generic use of the word to refer to any human being
  `date#1` day of the month
  `date#3` appointment, engagement

  They are generally placed high in the WordNet hierarchy and are related senses of highly polysemous words.

- *Stop Senses* synsets which appear frequently in different contexts, such as numbers, week days, colors, etc. These synsets usually belong to non polysemous words and they behaves much as *stop words*, because they do not significantly contribute to the overall sense of a text.

We have identified 2780 *stop senses* and 3670 *generics* in Wordnet 1.6, which results in 6450 synsets belonging to the FACTOTUM SFC.

3. Annotation procedure on WordNet 1.6

The procedure for the annotation of the synsets with SFCs iterates three steps. First, a small number of high level synsets are manually annotated with their pertinent SFCs. Then, an automatic procedure exploits some of the WordNet relations (i.e. hyponymy, troponymy, meronymy, antonymy and pertain-to) to extend the manual assignments to all the reachable synsets. As an example, this inheritance-based procedure allows to automatically mark the synset {beak, bill, neb, nib} with the code ZOOLOGY, starting from the synset {bird} and following a "part-of" relation.
There are cases in which the inheritance procedure has to be blocked, inserting an "exception", to prevent a wrong propagation. For instance, `barber_chair#1`, being a "part-of" `barbershop#1`, which in turn is annotated with COMMERCE, would wrongly inherit the same SFC. To deal with these cases, the inheritance procedure allows the declaration of exceptions, such as:

```
assign shop#1 to commerce
   with exception [part, isa, shop#1]
```

Subject field	*Annotations #*	*Relevance %*	*Productivity*	*Precision*
administration	1969	2.70	103	0.93
agriculture	248	0.34	8	1.00
alimentation	2563	3.52	170	1.00
anthropology	298	0.40	9	0.97
archaeology	47	0.06	4	1.00
architecture	3151	4.33	10	0.95
art	2145	2.94	16	0.98
artisanship	49	0.06	12	1.00
astrology	16	0.02	4	1.00
astronomy	420	0.57	8	0.95
biology	20266	27.85	122	0.99
chemistry	2029	2.78	19	1.00
commerce	533	0.73	6	0.94
computer-science	452	0.62	4	1.00
earth	4035	5.54	39	1.00
economy	2393	3.28	6	0.98
engineering	563	0.77	26	1.00
fashion	748	1.02	26	1.00
history	1097	1.50	11	0.95
industry	768	1.05	7	0.91
law	1179	1.62	7	0.99
linguistics	1460	2.00	28	0.96
literature	619	0.85	15	0.98
mathematics	575	0.79	8	0.97
medicine	2660	3.65	15	0.99
military	1198	1.64	13	1.00
pedagogy	517	0.71	8	1.00
philosophy	162	0.22	6	0.99
physics	1387	1.90	9	0.98
play	456	0.62	17	1.00
politics	848	1.16	6	0.95
psychology	1714	2.35	22	0.70
publishing	348	0.47	16	0.95
religion	1707	2.34	7	1.00
sexuality	170	0.23	4	0.93
sociology	617	0.84	23	0.73
sport	2520	3.46	7	0.94
telecommunication	463	0.63	6	0.89
tourism	409	0.56	4	0.90
transport	1623	2.23	9	0.94
veterinary	36	0.04	6	1.00
factotum	8305	11.41	82	0.90
Total	72763	100	22.09	0.95

Table 1: SFC distribution in WN16 (noun taxonomy)

which assigns the synset `shop#1` to COMMERCE, but excludes all the parts of the children of `shop#1`, such as `barbershop#1`.

Finally, the results of the inheritance procedure are evaluated in a text classification task: the main problems are detected and the manual annotations are corrected, starting a new iteration of the mapping procedure.

The effort needed to built the resource is given by a combination of the *productivity* of a subject field with its *relevance* with respect to WordNet.

Productivity. This is the ratio between the number of manual annotations and the total number of annotated synsets after the inheritance step. From the annotator point of view, this measure indicates how "difficult" it has been to produce a certain SFC. It also gives an idea of the homogeneity of the SFC with respect to the WordNet hierarchy: a low score in productivity generally means that synsets are spread in many sub-hierarchies. For example, BIOLOGY has a very high productivity (i.e. 122), mainly because it includes scientific taxonomies. On the contrary, ECONOMY has lower productivity (i.e. 6), meaning that it takes senses from different WordNet areas.

Relevance. This measure indicates the relative coverage of the subject field, i.e. how large it is compared to the total sum of the annotated synsets. As an example, MEDICINE and ALIMENTATION have similar relevance (3.65% and

3.52% respectively), but different productivity (15 and 170 respectively), meaning that an higher effort is needed for building MEDICINE rather than ALIMENTATION.

3.1. SFCs and Regular Polysemy

Regular polysemy (Pustejovsky, 1995) occurs when two senses of the same words are systematically related by a logical relation. Common examples are the `garlic#1` and `garlic#2` senses, where the first is a plant and the second is the food derived from that plant. In Wordnet 1.6 there is no rigorous treatment of regular polysemy, that is sometimes the two senses are clearly distinguished, as in the example above; sometimes there is just one sense, which however inherits from two different paths, as for example `Venice#1` which is both a region and an administrative district; sometimes the two senses are completely collapsed as for example in `Paris#1`. As far as semantic fields are concerned, we decided to be the more insensible to WordNet idiosyncrasies as possible. As instance:

`hospital#1` $\longrightarrow$ [MEDICINE, BUILDING-INDUSTRY]
`hospital#2` $\longrightarrow$ [MEDICINE, ADMINISTRATION]
`railway-station#1` $\longrightarrow$ [RAILWAY, BUILDING-INDUSTRY, ADMINISTRATION]

4. Results

Up to now 96% of the WordNet noun synsets have been marked (i.e. about 63,000 out of 66,000). We plan to conclude the annotation in a short time. Verbs and adjectives are under development and will be completed in the near future. 115 different SFCs, organized in a four level hierarchy have been used for the annotation. Table 1 shows the main figures of the assignments for the 41 SFCs placed at the second level of the hierarchy, which we consider the most informative one, plus the FACTOTUM SFC (see Section 2.3.). For each SCF the total number of synset assigned, its relevance and productivity measure, are reported. In addition, we show the precision score obtained for each SFC in the evaluation experiment, which is described in detail in Section 5.. Globally, up to now 72,763 mappings synset/SFC have been established, with an average ambiguity rate of 1.09%; 95% synsets have just a single assignment, 3% have two assignments and 2%, mainly due to regular polysemy (see Section 3.1.), have three assignments.

5. Evaluation

A quality evaluation of the SFC annotation has been carried on by means of a text classification task. This is comparable to Walker and Amsler's work (Walker and Amsler, 1986), where they use the LDOCE subject codes to identify the subject matter of wire service stories taken from The New York Times. In our evaluation task we used a corpus of short news from Adnkronos (an example is reported in Figure 2), an important Italian news provider. News are the English translation of Italian news [2], and they mainly

[2]The availability of the original news in Italian let us open the possibility in the future to test the assumption that SFC are basically language independent.

CULTURE: GIOTTO- PAID BY MONKS TO WRITE ANTI-FRANCISCAN POETRY

Rome,10 Jan. -(Adnkronos)- Giotto was 'paid' to attack a faction of the Franciscans, the Spiritual ones, who opposed church decoration in honour of Poverello di Assisi.This has been revealed in the research of an Italian scholar who is a professor at Yale University, Stefano Ugo Baldassarri, who thinks he has solved the mystery of the only known poetry by the famous Tuscan painter: the Giotto verses have in fact always provoked wonder because they seem to be a criticism of the ideals of St. Francis and all the more so since their author was also the man who painted the famous frescoes of the Basilica at Assisi. ....

Figure 2: Sample text used for evaluation.

reports facts relevant for Italy; topics are extremely various, from medicine to politics, to sport and culture. The goal of the experiment was to obtain a measure of the quality of the SFC annotation performed by the semi-automatic procedure described in Section 3..

We started working on a training corpus of 100 news. Each news has been manually classified using a SFC selected from the list of the 41 most informative SFC (see Section 4.). Texts have been POS tagged considering only Nouns, Proper Nouns and Abbreviations.

A classification algorithm has been implemented which uses the SFC assignments to compute the probability that senses of a word in a text belong to a certain subject code. Then, for each SFC an overall score is computed taking into account the SFC hierarchy (i.e. children contribute to the score of their father). At the end, the SFC with higher score is selected as the best to classify a given text.

While at present the system performances as text classifier are not the major point, the classification errors give us cues about the quality of SFC annotations. In particular:

- an high percentage of wrong classifications on the same SFC is an hint either of a overgeneration of the inheritance procedure for that code (e.g. additionally exceptions are needed), or of a wrong position of the SFC in the hierarchy; this for example happened with SFC whose content was not enough clear, such as PSYCHOLOGY, which in fact has a low precision score (see Table 1).

- the inability to classify a code indicates a lack of assignments for that code (i.e. more low level synsets need to be added).

In a more advanced phase of evaluation, a more fine grained mechanism was needed. In this phase we used the training corpus to manually check the output of the classification algorithm (see Table 3): errors in SFC annotation were detected and then corrected in the declarations of

the inheritance procedure. For example, `criticism#2` in Table 3, seems to be too general to be included annotated in LITERATURE, and so it has to be reassigned. Incidentally, let us note that the word "mistery" in the text, would be wrongly disambiguated with `mistery#2`.

After a number of iterations on the training corpus, the final evaluation was carried out on a test corpus with the same characteristics. Results are reported in Table 2. The average polysemy has been computed with respect to WordNet 1.6, considering only Nouns, Proper Nouns and Abbreviations. Checked senses are those used by the classification algorithm.

	Training	*Test*
Total tokens	22,269	24,030
Tokens considered	7,446	7,841
Average polysemy	4.74	4.95
Checked senses	12,359	13,588
Precision	0.92	0.95
Recall	0.96	0.96

Table 2: Evaluation of the SFC assignments.

6. Conclusions and Future Work

Up to now 96% of the WordNet synsets of the noun hierarchy have been annotated. We have used 115 different SFCs, which are organized in a four level hierarchy. Annotating Wordnet noun synsets allow to group synsets which belong to the same domain, even if they belong to different sub-hierarchies.

Verbs and adjectives are under development and will be completed in the near future in order to provide cross-categorial information, which is almost absent in WordNet.

For each SFC, two measures (i.e. productivity and relevance) have been defined to quantify the effort needed to build the resource. In addition, to produce a qualitative evaluation of the SFC annotations, a text classification task has been carried out which uses SFCs as tags. Both precision and recall are satisfactory, and can be further improved with limited effort.

Even if at present the classification performances are out of our scopes, the system performed quite well, so that we intend to go deeper in this direction.

Another evaluation task we would like to carry out in the near future is a comparison of our SFCs against approaches (e.g. (Grefenstette, 1994), (Lin, 1998)), which produce sets of related word considering thesaurus-like information.

7. Acknowledgments

We would like to thank Adnkronos for having made available the news for the evaluation; Emanuele Pianta for the implementation of the mapping procedure, Lisa Zenoniani for her lexicographic work and Adam Killgarriff who offered helpful comments.

8. References

Agirre, E., G. Rigau, L. Padro, and J. Atserias, to appear. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities. Special Issue on SensEval*. To appear.

Artale, Alessandro, Anna Goy, Bernardo Magnini, Emanuele Pianta, and Carlo Strapparava, 1998. Coping with wordnet sense proliferation. In *First International Conference on Language Resources & Evaluation*. Granada, Spain.

Diekema, Anne, 1998. Dewey decimal classification. Http://istweb.syr.edu/ isdp561/Dewey/dui.html.

Dyvik, Helge, 1998. Translation as semantic mirrors. In *The 13th Biennal European Conference on Artificial intelligence*. ECAI'98.

Gonzalo, Julio, Felisa Verdejo, Carlos Peters, and Nicoletta Calzolari, 1998. Applying eurowordnet to cross-language text retrieval. *Computers and the Humanities*, 32(2-3):185–207.

Grefenstette, Gregory, 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Gythrie, Joe A., Louise Guthrie, Yorick Wilks, and Homa Aidinejad, 1991. Subject-dependent co-occurrence and word sense disambiguation. In *29th Annual meeting of the Association for Computational Linguistics*. Berkeley, California, USA.

Hearst, Marti A. and Hinrich Schutze, 1996. *Corpus Processing for Lexical Acquisition*, chapter Customizing a Lexicon to Better Suit a Computational Task. The MIT Press, pages 77–96.

Leacock, Claudia, Geoffrey Towell, and Ellen M. Voorhees, 1996. *Corpus Processing for Lexical Acquisition*, chapter Towards Building Contextual representaions of WordSenses Using Statistical Models. The MIT Press, pages 77–96.

Liddy, Elizabeth and Woojin Paik, 1993. Document filtering using semantic information from a machine readable dictionary. In *Workshop on Very Large Corpora*. ACL.

Lin, Dekang, 1998. Automatic retrieval and clusterind of similar words. volume II. Montreal: COLING-ACL'98.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller, 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Pustejovsky, James, 1995. *The Generative Lexicon*. Massachusetts Institute of Technology.

Schutze, Hinrich, 1992. Dimensions of meaning. minneapolis, MN: Supercomputing'92.

Schutze, Hinrich, 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Veronis, Jean and Nancy Ide, 1990. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th international Conference on Computational Linguistics*, volume 2. helsinki: COLING'90.

Walker, D. E. and R. A. Amsler, 1986. *Analyzing Language in Restricted Domain. Sublanguage description and Processing*, chapter The use of Machine Readable Dictio-

Sense description	*Partial score*	*Occurrences*
LITERATURE:		
criticism#3 n a written evaluation of a work of literature	0.33	1
verse#3 n a line of metrical text	1.0	2
argument#4 n a summary of the subject or plot of a literary wor ...	0.20	1
poetry#2 n any communication resembling poetry in beauty or t ...	1.00	4
verse#2 n a piece of poetry	1.0	2
journal#1 n a daily written record of (usually personal) exper ...	0.20	1
verse#1 n literature in metrical form	1.0	2
work#5 n the total output of a writer or artist (or a subst ...	0.14	2
figure#10 n language used in a figurative or nonliteral sense ...	0.07	1
library#2 n a collection of literary documents or records kept ...	0.20	2
mystery#2 n a story about a crime (usually murder) presented a ...	0.50	1
author#1 n writes (books or stories or articles or the like) ...	0.50	1
poetry#1 n literature in metrical form	1.0	4
codex#1 n an unbound manuscript of some ancient classic	1.00	1
criticism#2 n a serious examination and judgment of something; " ...	0.33	1
RELIGION:		
convent#2 n a community of people in a religious order (especi ...	1.00	1
chapel#2 n a service conducted in a chapel; "he was late for ...	1.00	1
monk#1 n a male religious living in a cloister and devoting ...	1.00	1
spiritual#1 n a kind of religious song originated by Blacks in t ...	1.00	1
convent#1 n a religious residence especially for nuns	1.00	1
church#1 n a group of Christians; any group professing Christ ...	1.00	2
Vatican#1 n the residence of the Catholic Pope in the Vatican ...	1.00	1
church#2 n for public (especially Christian) worship; "the ch ...	1.00	2
Franciscan#1 n a Catholic friar wearing the gray habit of the Fra ...	1.00	2
church#3 n a service conducted in a church; "don't be late fo ...	1.00	2
basilica#1 n an early Christian church designed like a Roman ba ...	0.50	1
chapel#1 n a place of worship that has its own altar	1.00	1
spiritualist#1 n someone who serves as an intermediary between the ...	1.00	1

Table 3: Sample text used for evaluation.

raries in Sublanguage Analysis. Hillsdale NJ Lawrence Earlbaum.

Wilks, Yorick and Mark Stevenson, 1998. Word sense disambiguation using optimised combination of knowledge sources. COLING-ACL'98.

Yarowsky, David, 1993. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*. Princeton, NJ.