Sense clusters for Information Retrieval: Evidence from Semcor and the EuroWordNet InterLingual Index

Julio Gonzalo, Irina Chugur and Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia (UNED)
{julio,irina,felisa}@ieec.uned.es

Abstract

We examine three different types of sense clustering criteria with an Information Retrieval application in mind: methods based on the wordnet structure (such as generalization, cousins, sisters...); co-occurrence of senses obtained from Semcor; and equivalent translations of senses in other languages via the EuroWordNet InterLingual Index (ILI). We conclude that a) different NLP applications demand not only different sense granularities but different (possibly overlapped) sense clusterings. b) co-occurrence of senses in Semcor provide strong evidence for Information Retrieval clusters, unlike methods based on wordnet structure and systematic polysemy. c) parallel polysemy in three or more languages via the ILI, besides providing sense clusters for MT and CLIR, is strongly correlated with co-occurring senses in Semcor, and thus can be useful for Information Retrieval as well.

1 Introduction

WordNet (Miller et al., 1990) and EuroWordNet (Vossen, 1998), as most large-coverage electronic dictionaries and semantic networks, are not designed for a specific Natural Language Processing (NLP) application. It is commonly assumed that sense distinctions in these lexical databases are too fine-grained for a majority of applications. In particular, we have used EuroWordNet in a Cross-Language Text Retrieval (CLTR) application (Verdejo et al., 2000) and a number of CLTR experiments (Gonzalo et al., 1999; Vossen et al., 1999), confirming that it is crucial to apply certain sense clusters to Wordnet (WN) and EuroWordNet (EWN) to take real advantage of them in Information Retrieval applications. Potentially, a semantic network such as WN/EWN can help distinguishing different word senses for retrieval, enhancing precision, and identifying synonymic or conceptually related terms, enhancing recall. But not all sense distinctions in a lexical database are meaningful for Information Retrieval. For instance, the following sense distinctions are superfluous in an information retrieval application, as the different senses do not lead to different topics or different kinds of documents:

Behaviour
1. Manner of acting or conducting oneself
2. (psychology) the aggregate of the responses or reaction or movements made by an organism in any situation
3. Behavioural attributes

Bet
1. The act of gambling
2. The money risked on a gamble

Band
8. Instrumentalists not including string players
9. A group of musicians playing popular music for dancing

Bother
1. Smth. or someone who causes trouble, a source of unhappiness
2. An angry disturbance

But sense clustering have been generally associated with identifying Systematic Polysemy rules, taking into account lexicographic arguments rather than potential applications. In (Peters et al., 1998; Peters and Peters, 2000) the EuroWordNet structure, together with systematic polysemy, is used to group senses (sisters, auto-hyponymy, cousins, twins). This work is linked to the finding of systematic polysemy classes in (Buitelaar, 1998; Buitelaar, 2000; Tomuro, 1998) and others.

While identifying systematic polysemy might be a key issue for and adequate lexico-semantic spec-

ification, systematic relatedness does not always mean sense proximity. In particular, such rules do not necessarily predict a similar behavior of the clustered senses in an NLP application. For instance, the *animal/food* systematic polysemy does not lead to good sense clusters neither for Machine Translation between English and Spanish, nor for Information Retrieval . In Spanish it is common to give different names to an animal in a zoological sense or in a food sense. For instance, it is necessary to distinguish animal/food senses of *fish* in order to translate into *pez* or *pescado*, depending on the context. And for Information Retrieval, the *animal* sense will appear in documents about, say, zoology, while the *food* sense will appear in documents about cooking. Therefore, while the animal/food rule is useful for lexical representation and prediction of sense extensions in English, it cannot be used to cluster senses in MT or IR.

In (Vossen et al., 1999) we performed a concept-based IR experiment where using the ILI with clusters was slightly worse than using the ILI without the clusters. While clustering the EWN Interlingual Index records on the basis of systematic polysemy proved useful to provide better inter-languages connectivity in the EWN database, this result supports the idea that systematic polysemy, per se, is not an indication of potential IR clusters.

However, we do not claim that all systematic polysemy patterns are useless for IR. It is probably reasonable to classify different systematic polysemy rules according to whether they produce IR clusters or not. Some, already identified, patterns of regular polysemy, such as container/quantity or music/dance (Peters and Peters, 2000) yield adequate IR clusters. Other patterns, such as animal/food, plant/food, animal/skin, language/people tend to produce clusters that are not valid for IR. This classification of polysemy patterns is, to our opinion, strongly related with the black and white dot operators introduced in (Buitelaar, 1998). The black operator was reserved for polysemy patterns including sets of senses that may co-occur in the same word instance (e.g. book as written work or as physical object), and white operator is reserved for polysemy patterns for senses that never co-occur in the same word instance (e.g. window as physical object or as computer frame). Unfortunately, the distinction between black and white operators classes has not been applied yet -to our knowledge - to the set of polysemous classes defined in Buitelaar's thesis.

But, in many cases, even useful polysemy rules fail to extract pairs of systematically related senses in WN/EWN, because the hypernym branches that they pertain to do not obey none of the described systematic polysemy classes/types. Take the following example:

sack:
1. The act of terminating someone's employment → TERMINATION, END, CONCLUSION
2. a bag made of paper or plastic for holding customer purchases → BAG
3. unwaisted loose-fitting dress hanging straight from the shoulders → DRESS, FROCK
4. hanging bed of canvas or rope netting → BED
5. a woman's full loose hip-length jacket → JACKET
6. dry white wine from SW Europe → WHITE WINE
7. quantity contained in a sack → CONTAINERFUL
8. pocket. → ENCLOSED SPACE

sack 2 (bag of paper for customer's purchases) and *sack 7* (quantity contained in a sack) are related by systematic polysemy as container/containerful. Similarly, *sack 8* (pocket) should be related to some sense with the meaning of quantity. Nevertheless, *sack 8*, whose hypernym is "*enclosed space*", cannot be retained in the same way that the former pair of senses, in spite of identical semantic relationship. Systematic polysemy cannot predict, as well, a potential IR cluster with senses 3 and 5 (both meaning types of clothing and thus likely to appear in similar contexts). Senses 3 and 5 indicate, also, that clustering might also depend on the application domain: they can be clustered in a generic search, but they should be distinguished if the search is performed in a clothing domain.

It is interesting to note, finally, that different clustering criteria not only lead to different granularities, but they can produce tangled clusters, as in

Onion:
1. Pungent bulb → VEGETABLE → FOOD
2. Bulbuos plant having hollow leaves cultivated worldwide for its rounded edible bulb → ALLIACEOUS PLANT → PLANT
3. Edible bulb of an onion plant → BULB → PLANT ORGAN

The plant/food rule successfully relates senses 2 and 1, while for Information Retrieval the interesting cluster is for senses 2 and 3, (both botanical terms).

Our hypothesis is, therefore, that we cannot assume general clustering criteria; different NLP applications require different clustering criteria that are difficult to reconcile in a single clustering approach. Our work on clustering is centered on identifying sense-distinctions that could be relevant from an Information Retrieval and Cross Language Information Retrieval point of view.

Next section describes a clustering strategy that adequates to the Information Retrieval criterion: cluster senses if they tend to co-occur in the same Semcor documents.

In Section 3, we study a different clustering criterion, based on equivalent translations for two or more senses in other wordnets from the EuroWordNet database. This is a direct criterion to cluster senses in Machine Translation or Cross-Language Text Retrieval. Then we measure the overlap between both criteria, to conclude that the EWN InterLingual Index is also a valuable source of evidence for Information Retrieval clusters.

2 Cluster evidence from Semcor

One of our goals within the EuroWordNet and ITEM projects was to provide sense clusterings for WordNet (and, in general,for the EuroWordNet InterLingual Index, (Gonzalo et al., 1999) that leave only the sense distinctions in wordnets that indicate different (semantic) indexing units for Information Retrieval. Our first lexicographic examination of WordNet sense distinctions and clusterings following criteria based on the wordnet hierarchy did not produce clear criteria to classify senses semi-automatically according to this IR requirement. As we mentioned before, the clusters applied on the EWN InterLingual Index which relied solely on hierarchical information in Wordnet, produced a slight decrease of retrieval performance in an experiment using ILI records as indexing units.

Thus, we decided to stick to our only clear-cut criterion: cluster senses if they are likely to co-occur in the same document. The fact that the same sense combination occurs in several semantically tagged documents should provide strong evidence for clustering. Fortunately, we had the Semcor corpus of semantically-tagged documents to start with.

For example, the first two senses of "breath" co-occur in several Semcor documents:

Breath
1. (the air that is inhaled or exhaled in respiration)
2. (the act of exhaling)

This co-occurrence indicates that this sense distinction will not help to discriminate different document contexts. While in this particular example there is a clear relation between senses (sense 1 is involved in the action specified in sense 2), it seems extremely difficult to find general clustering techniques based on WordNet hierarchy to capture all potential IR clusters.

We have scanned Semcor in search of sets of (two or more) senses that co-occur frequently enough. In practice, we started with a threshold of at least 2 documents (out of 171) with the co-occurring senses in a similar distribution. We did not use the original Semcor files, but the IR-Semcor partition (Gonzalo et al., 1998) that splits multi-text documents into coherent retrieval chunks. We completed this list of candidates to cluster with pairs of senses that only co-occur once but belong to any "cousin" combination (Peters et al., 1998). Finally, we obtained 507 sets of sense combinations (producing above 650 sense pairs) for which Semcor provides positive evidence for clustering. A manual verification of 50 of such clusters showed that above 70% of them were useful. We also noticed that raising the threshold (the number of documents in which the senses co-occur), the error rate decreases quickly.

Then we worked with this set of positive IR clusters, trying to identify a set of common features that could be used to cluster the rest of WN/EWN senses. However, it seems extremely difficult to find any single criterion, common to all clusters.

For instance, if we consider a) number of variants in common between the synsets corresponding to the candidate senses; b) number of words in common between the glosses; and c) common hypernyms, we find that any combination of values for these three features is likely to be found among the set of clusters inferred from Semcor. For example:

fact
1. a piece of information about circumstances that exist or events that have occurred; "first you must collect all the facts of the case"
2. a statement or assertion of verified information about something that is the case or has happened; "he supported his argument with an impressive array of facts"
Number of documents in which they co-occur: 13
a) number of variants in common: 1 out of 1
b) (content) words in common between glosses: yes
c) common hypernyms: no

door

1. door –(a swinging or sliding barrier that will close the entrance to a room or building; "he knocked on the door"; "he slammed the door as he left") 2. doorway, door, entree, entry, portal, room access – (the space in a wall through which you enter or leave a room or building; the space that a door can close; "he stuck his head in the doorway") Number of documents in which they co-occur: 11
a) number of variants in common: 1 out of 6
b) (content) words in common between glosses: yes (also XPOS: enter/entrance)
c) common hypernyms: yes

way

1. manner, mode, style, way, fashion – (a manner of performance; "a manner of living"; "a way of life") 2. means, way – (how a result is obtained or an end is achieved; "a means of communication"; "the true way to success") Number of documents in which they co-occur: 9
a) number of variants in common: 1 out of 6
b) (content) words in common between glosses: no
c) common hypernyms: no

The next logical step is to use this positive evidence, combined with negative co-occurrence evidence, in training some machine learning system that can successfully capture the regularities hidden to our manual inspection. In principle, a binary classification task would be easy to capture by decision trees or similar techniques.

Therefore, we have also extracted from Semcor combinations of senses that appear frequently enough in Semcor, but never appear together in the same document. The threshold was set in, at least, 8 occurrences of each sense in Semcor, resulting in more than 500 negative clusters. A manual verification of 50 of these negative clusters showed that about 80% of them were acceptable for Information Retrieval as senses that should be distinguished. Together with the positive evidence, we have more than 1100 training cases for a binary classifier. Our plan is to apply this classifier to the whole EWN InterLingual Index, and then perform precision/recall tests in the environment described in (Gonzalo et al., 1998; Gonzalo et al., 1999).

3 Cluster evidence from the ILI

When translated into a target language, sense distinctions of a word may be lexicalized. For instance, the English term *spring* is translated into Spanish as *primavera* in its "season" sense, into *muelle* in its "metal device" sense, or as *fuente* in its "fountain" sense. For an English-Spanish Machine Translation system, it is crucial to distinguish these three senses of *spring*. But it is also frequent that two or more senses of a word are translated into the same word, for one or more languages. For instance, child as "human offspring (son or daughter) of any age" and child as "young male person" are both translated into "niño" in Spanish, into "enfant" in French, and into "kind" in German. We will use the term "parallel polysemy" to refer to this situation in the rest of this article.

Obviously, a Machine Translation system does not need to distinguish these two senses. But it is also tempting to hypothesize that the existence of parallel polysemy in two or more target languages may indicate that the two senses are close enough to be clustered in more applications. Indeed, in (Resnik and Yarowsky, 1999) this criterion is proposed to determine which word senses should be retained or discarded in a testbed for automatic Word Sense Disambiguation systems.

In particular, our goal has been to test whether two or more senses of a word are likely to be clustered, for IR purposes, if they have parallel polysemy in a certain number of languages via the EuroWordNet *InterLingual Index*. If the answer is positive, then the InterLingual Index, with eight languages interconnected, would be a rich source of information to provide IR clusters. In EWN, each monolingual database is linked, via Cross-Language equivalence relations, to the InterLingual Index (ILI) which is the superset of all concepts occurring in all languages. The ILI permits finding equivalent synsets between any pair of languages included in the database. For instance, senses 1 and 2 of *child* are translated into Spanish, French and German as follows:

Child

child 1 → {*child, kid*} – *(a human offspring (son or daughter) of any age; "they had three children"; "they were able to send their kids to college")*
child 2 → {*male child, boy, child*} – *(a young male person; "the baby was a boy"; "she made the boy brush his teeth every night")*

Spanish:
{***child***, *kid*} EQ-SYNONYM {***niño***, *crío, menor*}

{*male child, boy,* ***child***} EQ-SYNONYM {***niño***}

French:
{***child***, *kid*} EQ-SYNONYM{***enfant***, *mineur*}
{*male child, boy,* ***child***} EQ-SYNONYM{***enfant***}

German:
{***child***, *kid*} EQ-SYNONYM {***kind***}
{*male child, boy,* ***child***} EQ-SYNONYM {***kind***, *spross*}

Note that *child 1* and *child 2* have parallel translations in all three languages: Spanish (*niño*), French (*enfant*) and German (*kind*). In this case, this criterion successfully detects a pair of senses that could be clustered for Information Retrieval purposes.

In order to test the general validity of this criterion, we have followed these steps:

- Select a set of nouns for a full manual study. We have chosen the set of 22 nouns used in the first SENSEVAL competition (Kilgarriff and Palmer, 2000). This set satisfied our requirements of size (small enough for an exhaustive manual revision), reasonable degree of polysemy, and unbiased for our testing purposes (the criteria to select these 22 nouns was obviously independent of our experiment). We had to reduce the original set to 20 nouns (corresponding to 73 EWN senses), as the other two nouns were polysemous in the Hector database used for SENSEVAL, but monosemous in WordNet 1.5 and EuroWordNet. As target languages we chose Spanish, French, Dutch and German.

- Extract the candidate senses that satisfy the parallel polysemy criterion, in three variants:
  - Experiment 1: sets of senses that have parallel translations in at least two out of the four target languages.
  - Experiment 2: sets of senses that have parallel translations in at least one out of the four target languages. This is a softer constraint that produces a superset of the sense clusters obtained in Experiment 1.
  - Experiment 3: sets of senses whose synsets are mapped into the same target synset for at least one of the target languages. This criterion cannot be tested on plain multilingual dictionaries, only on EWN-like semantic databases.

- Check out manually whether the clusters produced in Experiments 1-3 are valid for Information Retrieval. At this step, the validity of clusters was checked by a human judge. Unfortunately, we did not have the chance yet to attest the validity of these judgments using more judges and extracting inter-annotator agreement rates. We could compare annotations only on a small fraction of cases (15 sense pairs), which we use to make the criterion "valid for IR" precise enough for reliable annotation. The results are reported in sections 3.2-3.4 for the different experiments.

- Identify all possible lexicographic reasons behind a parallel polysemy, taking advantage of the previous study. This is reported in the next section.

- Check how many clusters obtained from Semcor also satisfy the parallel translation criterion, to have an idea of the overlap between both (section 3.5).

- Finally, study whether the results have a dependency on possible incompleteness or inadequacy of the InterLingual Index (section 3.6).

3.1 Typology of parallel polysemy

Parallel polysemy can also be a sign of some systematic relation between the senses. As it is said in (Seto, 1996), *"(..) There often is a one-to-one correspondence between different languages in their lexicalization behaviour towards metonymy, in other words, metonymically related word senses are often translated by the same word in other languages"*.

But the reasons for parallel polysemy are not limited only to systematic polysemy. In the case of the EWN database, we have distinguished the following causes:

1. There is a series of mechanisms of meaning extension, if not universal, at least, common to several languages:
   (a) **Generalization/specialization** For example, the following two senses for *band*:

   English: band*; French:* groupe*; German:* Band, Musicgruppe
   1. Instrumentalists not including string players
   2. A group of musicians playing popular music for dancing

Sense 1 is a specialization of Sense 2, and this pattern is repeated in French and German.

(b) **Metonymic relations**. Some of them form already well known systematic polysemy patterns. As for applicability to IR, we should be capable to discriminate regular polysemy rules that provide valid IR clusters from those that contain senses that can not be interpreted simultaneously within a same document. Examples include:

English: glass; Spanish: vaso
1. container
2. quantity

which is a valid IR cluster, and

English: rabbit; Spanish: conejo; French: lapin
1.mammal
2.meat

which should be distinguished for IR.

(c) **Metaphors**. This kind of semantic relation usually does not produce good IR clusters, because senses related by means of metaphor usually belong to different semantic fields and, consequently, tend to occur in distinct documents. For example:

English: giant; Spanish: coloso; French: colosse; Dutch: kolossus
1.a person of exceptional importance and reputation
2.someone who is abnormally large

(d) **Semantic calque or loan translation**. A (probably metaphorical) sense extension is copied in other languages. It also can produce undesirable clusters for IR, because the original relation between two senses involved can be based on a metaphor. For example:

English: window; Spanish: ventana; Dutch: venster.
1.an opening in the wall of a building to admit light and air
2.a rectangular part of a computer screen that is a display different of the rest of the screen
The original computer sense for *window* is also adopted in Spanish and German for the corresponding words *ventana* and *venster*.

2. In certain occasions, the particularities of how the wordnets have been built semi-automatically lead to a mimesis of the WN1.5 senses and, consequently, to parallel polysemy in several languages. These sense distinctions are not incorrect, but perhaps would be different if every monolingual wordnet had been constructed without WN 1.5 as a reference for semi-automatic extraction of semantic relations. An example:

Behaviour:
1. Manner of acting or conducting oneself (Spanish: comportamiento, conducta; French: comportement, conduite)
2. (psychology) the aggregate of the responses or reaction or movements made by an organism in any situation (Spanish: comportamiento, conducta; French: comportement)
3. Behavioural attributes (Spanish: comportamiento, conducta; French: comportement)

The question is what classes of parallel polysemy are dominant in EWN, and then whether parallel polysemy can be taken as a strong indication of a potential IR cluster. A preliminary answer to this question is reported in the next sections.

3.2 Experiment 1

Here we selected all sense combinations, in our 20 English nouns test set, that had parallel translations in at least two of the four target languages considered (Spanish, French, Dutch and German). We found 10 clusters: 6 were appropriate for Information Retrieval, 3 were judged inappropriate, and one was due to an error in the database:

Valid IR clusters

Band 1,2: something elongated, worn around the body or one of the limbs / a strip or stripe of a contrasting color or material (mapped into two different synsets in Spanish and French)

band 2,5: a strip or stripe of a contrasting color or material /a stripe of a contrasting color (mapped into different synsets in Spanish and French; only one translation into Dutch.)

band 8,9: instrumentalists not including string players / a group of musicians playing popular

music for dancing (linked to the same synset in German and in Dutch)

behaviour 1,2,3: manner of acting or conducting oneself /(psychology) the aggregate of the responses or reaction or movements made by an organism in any situation / behavioural attributes (two senses are sisters, and in general the distinction is not easy to understand; in two cases the Dutch synset is the same, and there is no Dutch translation for the other. In Spanish there are three synsets that mimic the English ones).

Bet 1,2: act of gambling/money risked (metonymy relation, translated into different synsets in Spanish and French. One or both translations missing for the other languages)

excess 3,4: surplusage / overabundance (different synsets in Spanish and French, one or both translations missing in the other languages).

inappropriate clusters

giant 5,6: a person of exceptional importance / someone who is abnormally large (metaphoric relation; linked to the same synset in Dutch, and to different synsets in Spanish and French)

giant 5,7: a person of exceptional importance / a very large person (metaphoric relation; linked to different synsets in Dutch and German)

rabbit 1,2: mammal / meat (systematic polysemy; linked to different synsets in Spanish, German and French).

Erroneous cluster

steer 1,2: castrated bull/ hint, indication of potential opportunity. Both are translated into "buey" in Spanish and into "stierkalf" in Dutch. Only the "castrated bull" → "buey" link is appropriate.

3.3 Experiment 2

If we take all clusters that have a parallel translation in at least one target language (rather than two target languages as in Experiment 1), we obtain a larger subset of 27 clusters. The 17 new clusters have the following distribution:

- 9 valid clusters, such as *bother 1,2* (something that causes trouble / angry disturbance).
- 3 inappropriate clusters that relate homonyms, such as *band 2,7* (strip or stripe of a contrasting color or material/unofficial association of people).
- 4 inappropriate clusters that group metonymically related senses, such as *sanction 2,3* (penalty/authorization).
- 1 inappropriate cluster based on a metaphor: *steering 2,3* (act of steering and holding the course/guiding, guidance)

On the overall, we have 15 valid clusters, 11 inappropriate, and one error. The percentage of useful predictions is 56%, only slightly worse than for the tighter constraint of experiment 1. It is worth noticing that:

1. The parallel translation criterion obtained 27 potential clusters for 20 nouns, nearly one and a half cluster per noun. The criterion is very productive!
2. The percentage of incorrect clusters (41%) is high enough to suggest that parallel polysemy cannot be taken as a golden rule to cluster close senses, at least with the languages studied. Even 3 of the negative cases were homonyms, totally unrelated senses. Perhaps the general WSD clustering criterion proposed in (Resnik and Yarowsky, 1999) needs to be revised for a specific application such as IR. For instance, they argue that clusters based on parallel polysemy "would eliminate many distinctions that are arguably better treated as regular polysemy". But we have seen that regular polysemy may lead to sense distinctions that are important to keep in an Information Retrieval application. On the other hand, the results reported in (Resnik and Yarowsky, 1999) suggest that we would obtain better clusters if the parallel polysemy criteria is tested on more distant languages, such as Japanese or Basque to test English sense distinctions.

3.4 Experiment 3

In this experiment, which cannot be done with a multilingual dictionary, we looked for sense distinctions that are translated into *the same* synset for some target language. This is a direct evidence of sense relatedness (both senses point to the same concept in the target language), although the relation may be complex (for instance, one of the two senses might be translated as an EQ-HYPONYM).

We found 9 clusters satisfying the criterion, all of them for links to the Dutch wordnet. 5 sense combinations are valid IR clusters. Three combinations turned out to be inappropriate for the

# words = 20 # senses = 73	IR clusters	not IR clusters	incorrect	Total
Exp. 1	6 (60%)	3 (30%)	1	10
Exp. 2	15 (56%)	11 (41%)	1	27
Exp. 3	5 (56%)	3 (33%)	1	9

Table 1: Adequacy of clusters based on parallel polysemy for Information Retrieval

needs of IR (accident 1,2: chance / misfortune; steering 2,3: the act of steering and holding the course / guiding, guidance; giant 5,6: a person of exceptional importance / someone who is abnormally large). Finally, the erroneous cluster for steer1 (castrated bull) and steer2 (hint, an indication of potential opportunity) reappeared again.

The results for the three experiments are summarized in Table 1. It seems that the parallel polysemy criteria on the ILI can be a very rich source of information to cluster senses for IR, but it is as well obvious that it needs to be refined or manually revised to obtain high quality clusters.

3.5 Overlapping of criteria from Semcor to ILI

To complete evidence for correlation between Semcor-based clusters and ILI-based clusters, we studied two subsets of Semcor-based clusters to check if they matched the parallel polysemy criteria on the ILI. The first set were the 11 sense combinations with a co-occurrence frequency above 7 in Semcor. 10 out of 11 (91%) also hold the most restrictive criterion used in Experiment 1, again indicating a strong correlation between both criteria. Then we augmented the set of sense combinations to 50 - with co-occurrence frequencies above 2-. This time, 27 clusters matched the criterion in Experiment 2 (54%). As the evidence for Semcor clustering decreases, the criterion of parallel translations is also less reliable, again indicating a correlation between both.

3.6 Adequacy of the ILI to get translation clusters

Clustering methods based on the criterion of parallel translation depend, to a great extent, on the adequacy and quality of the lexical resources used. How many ILI clusters had we obtained in an EWN database with total coverage and completely error-free?

Our experiments, though limited, are a first indication of the utility of EWN for this task:

- Analyzing 73 WN senses corresponding to 20 nouns used in the SENSEVAL, we found 2 erroneous equivalence links in the Spanish and Dutch wordnets. Taking into account that EWN was built by semi-automatic means, this seems a low error rate.

- Only 16 senses out of 73 have equivalence links in the 4 selected wordnets. 19 senses have equivalence links in 3 languages, 21 senses in 2 languages, 9 in only one language and 6 have no equivalence links in any of the selected wordnets. The lack of equivalence links sometimes can be explained by the lack of lexicalized terms for a certain WN concept. For example, float2 (a drink with ice-cream floating in it) is not lexicalized in Spanish, so we should not expect an equivalence link for this sense in the Spanish wordnet. In many other cases though, the lack of the equivalence links is due to incompleteness in the database.

- Each monolingual wordnet reflects, to a large extent, the kind of Machine-Readable resources used to build it. The Spanish wordnet was built mainly from bilingual dictionaries and therefore is closer to the Wn 1.5 structure. The French wordnet departed from an ontology-like database, and thus some non-lexicalized expressions are still present (for instance, *float 2* has *soda_avec_un_boule_de_glace* as French equivalent). The Dutch wordnet departed from a lexical database rich in semantic information, thus it departs more from the Wordnet structure, has a richer connectivity and complex links into the InterLingual Index, etc. Cross-Language equivalent relations are not, therefore, totally homogeneous in EWN.

On the overall, however, the ILI seems perfectly suitable for automatic applications regarding multilingual sense mappings. In particular, the fine-grainedness of Wordnet and EuroWordNet, in spite of its lack of popularity among NLP researchers, may be an advantage for NLP applications, as it may suit different clusterings for different application requirements.

4 Conclusions

We examined three different types of sense clustering criteria with an Information Retrieval application in mind: methods based on the wordnet structure (such as generalization, cousins, sisters...); co-occurrence of senses obtained from Semcor; and equivalent translations of senses in other languages via the EuroWordNet InterLingual Index (ILI). We conclude that a) different NLP applications demand not only different sense granularities but different (possibly overlapped) sense clusterings. b) co-occurrence of senses in Semcor provide strong evidence for Information Retrieval clusters, unlike methods based on wordnet structure and systematic polysemy. c) parallel polysemy in two or more languages via the ILI, besides providing sense clusters for MT and CLIR, is correlated with coocurring senses in Semcor, and thus can be useful to obtain IR clusters as well.

Both approaches to IR clusters for WN/EWN (evidence from Semcor and from the ILI) seem very promising. The positive and negative evidence from Semcor (above 500 clusters each) can possibly be used in a Machine Learning approach to find additional clusters for the remaining sense distinctions without enough evidence from Semcor. The parallel polysemy criteria, over EWN, is highly productive (more than one candidate per word in our experiments), although a more diverse set of languages would probably produce a higher rate of valid clusters.

References

P. Buitelaar. 1998. *CoreLex: systematic polysemy and underspecification.* Ph.D. thesis, Department of Computer Science, Brandeis University, Boston.

P. Buitelaar. 2000. Reducing lexical semantic complexity with systematic polysemous classes and underspecification. In *Proceedings of ANLP'2000.*

J. Gonzalo, M. F. Verdejo, I. Chugur, and J. Cigarrán. 1.998. Indexing with Wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems.*

J. Gonzalo, A. Peñas, and F. Verdejo. 1999. Lexical ambiguity and information retrieval revisited. In *Proceedings of EMNLP/VLC'99 Conference.*

A. Kilgarriff and M. Palmer. 2000. Special issue on senseval. *Computers and the Humanities*, 34(1-2).

G. Miller, C. Beckwith, D. Fellbaum, D. Gross, and K. Miller. 1990. Five papers on Wordnet, CSL report 43. Technical report, Cognitive Science Laboratory, Princeton University.

W. Peters and I. Peters. 2000. Automatic sense clustering in EuroWordnet. In *Proceedings of LREC'2000.*

W. Peters, I. Peters, and P. Vossen. 1998. Lexicalized systematic polysemy in EuroWordNet. In *Proceedings of the First International Conference on Language Resources and Evaluation.*

P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering.*

K. Seto. 1996. On the cognitive triangle: the relation of metaphor, metonymy and synecdoque. In A. Burkhardt and N. Norrich, editors, *Tropic Truth.* De Gruyter.

N. Tomuro. 1998. Semi-automatic induction of systematic polysemy from wordnet. In *Proceedings of COLING/ACL'98 workshop on the use of wordnet in NLP applications.*

F. Verdejo, J. Gonzalo, A. Peñas, F. López, and D. Fernández. 2000. Evaluating wordnets in cross-language text retrieval: the item multilingual search engine. In *Proceedings LREC'2000.*

P. Vossen, W. Peters, and J. Gonzalo. 1999. Towards a universal index of meaning. In *Proceedings of SIGLEX'99.*

P. Vossen. 1998. *EuroWordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers.