Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance

Roberto Navigli
Dipartimento di Informatica
Università di Roma "La Sapienza"
Roma, Italy
navigli@di.uniroma1.it

Abstract

Fine-grained sense distinctions are one of the major obstacles to successful Word Sense Disambiguation. In this paper, we present a method for reducing the granularity of the WordNet sense inventory based on the mapping to a manually crafted dictionary encoding sense hierarchies, namely the Oxford Dictionary of English. We assess the quality of the mapping and the induced clustering, and evaluate the performance of coarse WSD systems in the Senseval-3 English all-words task.

1 Introduction

Word Sense Disambiguation (WSD) is undoubtedly one of the hardest tasks in the field of Natural Language Processing. Even though some recent studies report benefits in the use of WSD in specific applications (e.g. Vickrey et al. (2005) and Stokoe (2005)), the present performance of the best ranking WSD systems does not provide a sufficient degree of accuracy to enable real-world, language-aware applications.

Most of the disambiguation approaches adopt the WordNet dictionary (Fellbaum, 1998) as a sense inventory, thanks to its free availability, wide coverage, and existence of a number of standard test sets based on it. Unfortunately, WordNet is a fine-grained resource, encoding sense distinctions that are often difficult to recognize even for human annotators (Edmonds and Kilgariff, 1998).

Recent estimations of the inter-annotator agreement when using the WordNet inventory report figures of 72.5% agreement in the preparation of the English all-words test set at Senseval-3 (Snyder and Palmer, 2004) and 67.3% on the Open Mind Word Expert annotation exercise (Chklovski and Mihalcea, 2002). These numbers lead us to believe that a credible upper bound for unrestricted fine-grained WSD is around 70%, a figure that state-of-the-art automatic systems find it difficult to outperform. Furthermore, even if a system were able to exceed such an upper bound, it would be unclear how to interpret such a result.

It seems therefore that the major obstacle to effective WSD is the fine granularity of the WordNet sense inventory, rather than the performance of the best disambiguation systems. Interestingly, Ng et al. (1999) show that, when a coarse-grained sense inventory is adopted, the increase in inter-annotator agreement is much higher than the reduction of the polysemy degree.

Following these observations, the main question that we tackle in this paper is: *can we produce and evaluate coarse-grained sense distinctions and show that they help boost disambiguation on standard test sets?* We believe that this is a crucial research topic in the field of WSD, that could potentially benefit several application areas.

The contribution of this paper is two-fold. First, we provide a wide-coverage method for clustering WordNet senses via a mapping to a coarse-grained sense inventory, namely the Oxford Dictionary of English (Soanes and Stevenson, 2003) (Section 2). We show that this method is well-founded and accurate with respect to manually-made clusterings (Section 3). Second, we evaluate the performance of WSD systems when using coarse-grained sense inventories (Section 4). We conclude the paper with an account of related work (Section 5), and some final remarks (Section 6).

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 105–112,
Sydney, July 2006.

2 Producing a Coarse-Grained Sense Inventory

In this section, we present an approach to the automatic construction of a coarse-grained sense inventory based on the mapping of WordNet senses to coarse senses in the Oxford Dictionary of English. In section 2.1, we introduce the two dictionaries, in Section 2.2 we illustrate the creation of sense descriptions from both resources, while in Section 2.3 we describe a lexical and a semantic method for mapping sense descriptions of WordNet senses to ODE coarse entries.

2.1 The Dictionaries

WordNet (Fellbaum, 1998) is a computational lexicon of English which encodes concepts as synonym sets (*synsets*), according to psycholinguistic principles. For each word sense, WordNet provides a gloss (i.e. a textual definition) and a set of relations such as hypernymy (e.g. apple *kind-of* edible fruit), meronymy (e.g. computer *has-part* CPU), etc.

The Oxford Dictionary of English (ODE) (Soanes and Stevenson, 2003)[1] provides a hierarchical structure of senses, distinguishing between homonymy (i.e. completely distinct senses, like race as a competition and race as a taxonomic group) and polysemy (e.g. race as a channel and as a current). Each polysemous sense is further divided into a *core sense* and a set of *subsenses*. For each sense (both core and subsenses), the ODE provides a textual definition, and possibly hypernyms and domain labels. Excluding monosemous senses, the ODE has an average number of 2.56 senses per word compared to the average polysemy of 3.21 in WordNet on the same words (with peaks for verbs of 2.73 and 3.75 senses, respectively).

In Table 1 we show an excerpt of the sense inventories of the noun *race* as provided by both dictionaries[2]. The ODE identifies 3 homonyms and 3 polysemous senses for the first homonym, while WordNet encodes a flat list of 6 senses, some of which strongly related (e.g. $race\#1$ and $race\#3$). Also, the ODE provides a sense (ginger root) which is not taken into account in WordNet.

The structure of the ODE senses is clearly hierarchical: if we were able to map with a high accuracy WordNet senses to ODE entries, then a sense clustering could be trivially induced from the mapping. As a result, the granularity of the WordNet inventory would be drastically reduced. Furthermore, disregarding errors, the clustering would be well-founded, as the ODE sense groupings were manually crafted by expert lexicographers. In the next section we illustrate a general way of constructing sense descriptions that we use for determining a complete, automatic mapping between the two dictionaries.

2.2 Constructing Sense Descriptions

For each word w, and for each sense S of w in a given dictionary $D \in \{\text{WORDNET}, \text{ODE}\}$, we construct a sense description $d_{\text{D}}(S)$ as a bag of words:

$$d_{\text{D}}(S) = def_{\text{D}}(S) \cup hyper_{\text{D}}(S) \cup domains_{\text{D}}(S)$$

where:

- $def_{\text{D}}(S)$ is the set of words in the textual definition of S (excluding usage examples), automatically lemmatized and part-of-speech tagged with the RASP statistical parser (Briscoe and Carroll, 2002);
- $hyper_{\text{D}}(S)$ is the set of direct hypernyms of S in the taxonomy hierarchy of D ($\emptyset$ if hypernymy is not available);
- $domains_{\text{D}}(S)$ includes the set of domain labels possibly assigned to sense S ($\emptyset$ when no domain is assigned).

Specifically, in the case of WordNet, we generate $def_{\text{WN}}(S)$ from the gloss of S, $hyper_{\text{WN}}(S)$ from the noun and verb taxonomy, and $domains_{\text{WN}}(S)$ from the subject field codes, i.e. domain labels produced semi-automatically by Magnini and Cavaglià (2000) for each WordNet synset (we exclude the general-purpose label, called FACTOTUM).

For example, for the first WordNet sense of $race\#n$ we obtain the following description:

$$d_{\text{WN}}(race\#n\#1) = \{competition\#n\} \cup \{contest\#n\} \cup \{\text{POLITICS\#N}, \text{SPORT\#N}\}$$

In the case of the ODE, $def_{\text{ODE}}(S)$ is generated from the definitions of the core sense and the subsenses of the entry S. Hypernymy (for nouns only) and domain labels, when available, are included in the respective sets $hyper_{\text{ODE}}(S)$

[1] The ODE was kindly made available by Ken Litkowski (CL Research) in the context of a license agreement.

[2] In the following, we denote a WordNet sense with the convention *w#p#i* where w is a word, p a part of speech and i is a sense number; analogously, we denote an ODE sense with the convention $w\#p\#h.k$ where h is the homonym number and k is the k-th polysemous entry under homonym h.

Table 1: The sense inventory of *race#n* in WordNet and ODE (definitions are abridged, bullets (•) indicate a subsense in the ODE, arrows (→) indicate hypernymy, DOMAIN LABELS are in small caps).

race#n (WordNet)	
#1	Any competition (→ contest).
#2	People who are believed to belong to the same genetic stock (→ group).
#3	A contest of speed (→ contest).
#4	The flow of air that is driven backwards by an aircraft propeller (→ flow).
#5	A taxonomic group that is a division of a species; usually arises as a consequence of geographical isolation within a species (→ taxonomic group).
#6	A canal for a current of water (→ canal).

race#n (ODE)	
#1.1	**Core:** SPORT A competition between runners, horses, vehicles, etc. • RACING A series of such competitions for horses or dogs • A situation in which individuals or groups compete (→ contest) • ASTRONOMY The course of the sun or moon through the heavens (→ trajectory).
#1.2	**Core:** NAUTICAL A strong or rapid current (→ flow).
#1.3	**Core:** A groove, channel, or passage. • MECHANICS A water channel • Smooth groove or guide for balls (→ indentation, conduit) • FARMING Fenced passageway in a stockyard (→ route) • TEXTILES The channel along which the shuttle moves.
#2.1	**Core:** ANTHROPOLOGY Division of humankind (→ ethnic group). • The condition of belonging to a racial division or group • A group of people sharing the same culture, history, language • BIOLOGY A group of people descended from a common ancestor.
#3.1	**Core:** BOTANY, FOOD A ginger root (→ plant part).

and $domains_{\text{ODE}}(S)$. For example, the first ODE sense of $race\#n$ is described as follows:

$$d_{\text{ODE}}(race\#n\#1.1) = \{competition\#n, runner\#n, horse\#n, vehicle\#n, \ldots, heavens\#n\} \cup \{contest\#n, trajectory\#n\} \cup \{\text{SPORT\#N}, \text{RACING\#N}, \text{ASTRONOMY\#N}\}$$

Notice that, for every S, $d_{\text{D}}(S)$ is non-empty as a definition is always provided by both dictionaries. This approach to sense descriptions is general enough to be applicable to any other dictionary with similar characteristics (e.g. the Longman Dictionary of Contemporary English in place of ODE).

2.3 Mapping Word Senses

In order to produce a coarse-grained version of the WordNet inventory, we aim at defining an automatic mapping between WordNet and ODE, i.e. a function $\mu : Senses_{\text{WN}} \rightarrow Senses_{\text{ODE}} \cup \{\epsilon\}$, where $Senses_{\text{D}}$ is the set of senses in the dictionary D and ϵ is a special element assigned when no plausible option is available for mapping (e.g. when the ODE encodes no entry corresponding to a WordNet sense).

Given a WordNet sense $S \in Senses_{\text{WN}}(w)$ we define $\hat{m}(S)$, the best matching sense in the ODE, as:

$$\hat{m}(S) = \underset{S' \in Senses_{\text{ODE}}(w)}{\arg\max} \; match(S, S')$$

where $match : Senses_{\text{WN}} \times Senses_{\text{ODE}} \rightarrow [0, 1]$ is a function that measures the degree of matching between the sense descriptions of S and S'. We define the mapping μ as:

$$\mu(S) = \begin{cases} \hat{m}(S) & \text{if } match(S, \hat{m}(S)) \geq \theta \\ \epsilon & otherwise \end{cases}$$

where θ is a threshold below which a matching between sense descriptions is considered unreliable. Finally, we define the clustering of senses $c(w)$ of a word w as:

$$c(w) = \{\mu^{-1}(S') : S' \in Senses_{\text{ODE}}(w), \mu^{-1}(S') \neq \emptyset\} \cup \{\{S\} : S \in Senses_{\text{WN}}(w), \mu(S) = \epsilon\}$$

where $\mu^{-1}(S')$ is the group of WordNet senses mapped to the same sense S' of the ODE, while the second set includes singletons of WordNet senses for which no mapping can be provided according to the definition of μ.

For example, an ideal mapping between entries in Table 1 would be as follows:

μ(race#n#1) = race#n#1.1, μ(race#n#2) = race#n#2.1,
μ(race#n#3) = race#n#1.1, μ(race#n#5) = race#n#2.1,
μ(race#n#4) = race#n#1.2, μ(race#n#6) = race#n#1.3,

resulting in the following clustering:

$$c(race\#n) = \{\{race\#n\#1, race\#n\#3\}, \{race\#n\#2, race\#n\#5\}, \{race\#n\#4\}, \{race\#n\#6\}\}$$

In Sections 2.3.1 and 2.3.2 we describe two different choices for the $match$ function, respectively based on the use of lexical and semantic information.

2.3.1 Lexical matching

As a first approach, we adopted a purely lexical matching function based on the notion of lexical overlap (Lesk, 1986). The function counts the number of lemmas that two sense descriptions of a word have in common (we neglect parts of speech), and is normalized by the minimum of the two description lengths:

$$match_{LESK}(S, S') = \frac{|d_{\text{WN}}(S) \cap d_{\text{ODE}}(S')|}{\min\{|d_{\text{WN}}(S)|, |d_{\text{ODE}}(S')|\}}$$

where $S \in Senses_{\text{WN}}(w)$ and $S' \in Senses_{\text{ODE}}(w)$. For instance:

$$match_{LESK}(race\#n\#1, race\#n\#1.1) = \frac{3}{\min\{4,20\}} = \frac{3}{4} = 0.75$$
$$match_{LESK}(race\#n\#2, race\#n\#1.1) = \frac{1}{8} = 0.125$$

Notice that unrelated senses can get a positive score because of an overlap of the sense descriptions. In the example, $group\#n$, the hypernym of $race\#n\#2$, is also present in the definition of $race\#n\#1.1$.

2.3.2 Semantic matching

Unfortunately, the very same concept can be defined with entirely different words. To match definitions in a semantic manner we adopted a knowledge-based Word Sense Disambiguation algorithm, Structural Semantic Interconnections (SSI, Navigli and Velardi (2004)).

SSI[3] exploits an extensive lexical knowledge base, built upon the WordNet lexicon and enriched with collocation information representing semantic relatedness between sense pairs. Collocations are acquired from existing resources (like the Oxford Collocations, the Longman Language Activator, collocation web sites, etc.). Each collocation is mapped to the WordNet sense inventory in a semi-automatic manner and transformed into a *relatedness* edge (Navigli and Velardi, 2005).

Given a word context $C = \{w_1, ..., w_n\}$, SSI builds a graph $G = (V, E)$ such that $V = \bigcup_{i=1}^{n} Senses_{\text{WN}}(w_i)$ and $(S, S') \in E$ if there is at least one semantic interconnection between S and S' in the lexical knowledge base. A *semantic interconnection pattern* is a relevant sequence of edges selected according to a manually-created context-free grammar, i.e. a path connecting a pair of word senses, possibly including a number of intermediate concepts. The grammar consists of a small number of rules, inspired by the notion of lexical chains (Morris and Hirst, 1991).

SSI performs disambiguation in an iterative fashion, by maintaining a set $\mathcal{C}$ of senses as a semantic context. Initially, $\mathcal{C} = V$ (the entire set of senses of words in C). At each step, for each sense S in C, the algorithm calculates a score of the degree of connectivity between S and the other senses in $\mathcal{C}$:

$$Score_{SSI}(S, \mathcal{C}) = \frac{\sum_{S' \in \mathcal{C} \setminus \{S\}} \sum_{i \in IC(S,S')} \frac{1}{length(i)}}{\sum_{S' \in \mathcal{C} \setminus \{S\}} |IC(S,S')|}$$

where $IC(S, S')$ is the set of interconnections between senses S and S'. The contribution of a single interconnection is given by the reciprocal of its length, calculated as the number of edges connecting its ends. The overall degree of connectivity is then normalized by the number of contributing interconnections. The highest ranking sense S of word w is chosen and the senses of w are removed from the semantic context $\mathcal{C}$. The algorithm terminates when either $\mathcal{C} = \emptyset$ or there is no sense such that its score exceeds a fixed threshold.

Given a word w, semantic matching is performed in two steps. First, for each dictionary $D \in \{\text{WORDNET}, \text{ODE}\}$, and for each sense $S \in Senses_{\text{D}}(w)$, the sense description of S is disambiguated by applying SSI to $d_D(S)$. As a result, we obtain a semantic description as a bag of concepts $d_D^{sem}(S)$. Notice that sense descriptions from both dictionaries are disambiguated with respect to the WordNet sense inventory.

Second, given a WordNet sense $S \in Senses_{\text{WN}}(w)$ and an ODE sense $S' \in Senses_{\text{ODE}}(w)$, we define $match_{SSI}(S, S')$ as a function of the direct relations connecting senses in $d_{\text{WN}}^{sem}(S)$ and $d_{\text{ODE}}^{sem}(S')$:

$$match_{SSI}(S, S') = \frac{|c \rightarrow c' : c \in d_{\text{WN}}^{sem}(S), c' \in d_{\text{ODE}}^{sem}(S')|}{|d_{\text{WN}}^{sem}(S)| \cdot |d_{\text{ODE}}^{sem}(S')|}$$

where $c \rightarrow c'$ denotes the existence of a relation edge in the lexical knowledge base between a concept c in the description of S and a concept c' in the description of S'. Edges include the WordNet relation set (synonymy, hypernymy, meronymy, antonymy, similarity, nominalization, etc.) and the *relatedness* edge mentioned above (we adopt only direct relations to maintain a high precision).

For example, some of the relations found between concepts in $d_{\text{WN}}^{sem}(race\#n\#3)$ and $d_{\text{ODE}}^{sem}(race\#n\#1.1)$ are:

$race\#n\#3$	$relation$	$race\#n\#1.1$
speed#n#1	$\xrightarrow{related-to}$	vehicle#n#1
race#n#3	$\xrightarrow{related-to}$	compete#v#1
racing#n#1	$\xrightarrow{kind-of}$	sport#n#1
race#n#3	$\xrightarrow{kind-of}$	contest#n#1

contributing to the final value of the function on the two senses:

$$match_{SSI}(race\#n\#3, race\#n\#1.1) = 0.41$$

Due to the normalization factor in the denominator, these values are generally low, but unrelated

[3] Available online from: http://lcl.di.uniroma1.it/ssi

Table 2: Performance of the lexical and semantic mapping functions.

Func.	Prec.	Recall	F1	Acc.
Lesk	84.74%	65.43%	73.84%	66.08%
SSI	86.87%	79.67%	83.11%	77.94%

senses have values much closer to 0. We chose SSI for the semantic matching function as it has the best performance among untrained systems on unconstrained WSD (cf. Section 4.1).

3 Evaluating the Clustering

We evaluated the accuracy of the mapping produced with the lexical and semantic methods described in Sections 2.3.1 and 2.3.2, respectively. We produced a gold-standard data set by manually mapping 5,077 WordNet senses of 763 randomly-selected words to the respective ODE entries (distributed as follows: 466 nouns, 231 verbs, 50 adjectives, 16 adverbs). The data set was created by two annotators and included only polysemous words. These words had 2,600 senses in the ODE.

Overall, 4,599 out of the 5,077 WordNet senses had a corresponding sense in ODE (i.e. the ODE covered 90.58% of the WordNet senses in the data set), while 2,053 out of the 2,600 ODE senses had an analogous entry in WordNet (i.e. WordNet covered 78.69% of the ODE senses). The WordNet clustering induced by the manual mapping was 49.85% of the original size and the average degree of polysemy decreased from 6.65 to 3.32.

The reliability of our data set is substantiated by a quantitative assessment: 548 WordNet senses of 60 words were mapped to ODE entries by both annotators, with a pairwise mapping agreement of 92.7%. The average Cohen's κ agreement between the two annotators was 0.874.

In Table 2 we report the precision and recall of the lexical and semantic functions in providing the appropriate association for the set of senses having a corresponding entry in ODE (i.e. excluding the cases where a sense ϵ was assigned by the manual annotators, cf. Section 2.3). We also report in the Table the accuracy of the two functions when we view the problem as a classification task: an automatic association is correct if it corresponds to the manual association provided by the annotators or if both assign no answer (equivalently, if both provide an ϵ label). All the differences between Lesk and SSI are statistically significant ($p < 0.01$).

As a second experiment, we used two information-theoretic measures, namely *entropy* and *purity* (Zhao and Karypis, 2004), to compare an automatic clustering $c(w)$ (i.e. the sense groups acquired for word w) with a manual clustering $\hat{c}(w)$. The entropy quantifies the distribution of the senses of a group over manually-defined groups, while the purity measures the extent to which a group contains senses primarily from one manual group.

Given a word w, and a sense group $G \in c(w)$, the entropy of G is defined as:

$$H(G) = -\frac{1}{log\ |\hat{c}(w)|} \sum_{\hat{G} \in \hat{c}(w)} \frac{|\hat{G} \cap G|}{|\hat{G}|} log(\frac{|\hat{G} \cap G|}{|\hat{G}|})$$

i.e., the entropy[4] of the distribution of senses of group G over the groups of the manual clustering $\hat{c}(w)$. The entropy of an entire clustering $c(w)$ is defined as:

$$Entropy(c(w)) = \sum_{G \in c(w)} \frac{|G|}{|Senses_{\text{WN}}(w)|} H(G)$$

that is, the entropy of each group weighted by its size. The purity of a sense group $G \in c(w)$ is defined as:

$$Pu(G) = \frac{1}{|G|} \max_{\hat{G} \in \hat{c}(w)} |\hat{G} \cap G|$$

i.e., the normalized size of the largest subset of G contained in a single group $\hat{G}$ of the manual clustering. The overall purity of a clustering is obtained as a weighted sum of the individual cluster purities:

$$Purity(c(w)) = \sum_{G \in c(w)} \frac{|G|}{|Senses_{\text{WN}}(w)|} Pu(G)$$

We calculated the entropy and purity of the clustering produced automatically with the lexical and the semantic method, when compared to the grouping induced by our manual mapping (ODE), and to the grouping manually produced for the English all-words task at Senseval-2 (3,499 senses of 403 nouns). We excluded from both gold standards words having a single cluster. The figures are shown in Table 3 (good entropy and purity values should be close to 0 and 1 respectively).

Table 3 shows that the quality of the clustering induced with a semantic function outperforms both lexical overlap and a random baseline. The baseline was computed averaging among 200 random clustering solutions for each word. Random

[4]Notice that we are comparing clusterings against the manual clustering (rather than viceversa), as otherwise a completely unclustered solution would result in 1.0 entropy and 0.0 purity.

Table 3: Comparison with gold standards.

Gold standard	Method	Entropy	Purity
ODE	Lesk	0.15	0.87
	SSI	0.11	0.87
	Baseline	0.28	0.67
Senseval	Lesk	0.17	0.71
	SSI	0.16	0.69
	Baseline	0.27	0.57

clusterings were the result of a random mapping function between WordNet and ODE senses. As expected, the automatic clusterings have a lower purity when compared to the Senseval-2 noun grouping as the granularity of the latter is much finer than ODE (entropy is only partially affected by this difference, indicating that we are producing larger groups). Indeed, our gold standard (ODE), when compared to the Senseval groupings, obtains a low purity as well (0.75) and an entropy of 0.13.

4 Evaluating Coarse-Grained WSD

The main reason for building a clustering of WordNet senses is to make Word Sense Disambiguation a feasible task, thus overcoming the obstacles that even humans encounter when annotating sentences with excessively fine-grained word senses.

As the semantic method outperformed the lexical overlap in the evaluations of previous Section, we decided to acquire a clustering on the entire WordNet sense inventory using this approach. As a result, we obtained a reduction of 33.54% in the number of entries (from 60,302 to 40,079 senses) and a decrease of the polysemy degree from 3.14 to 2.09. These figures exclude monosemous senses and derivatives in WordNet. As we are experimenting on an automatically-acquired clustering, all the figures are affected by the 22.06% error rate resulting from Table 2.

4.1 Experiments on Senseval-3

As a first experiment, we assessed the effect of the automatic sense clustering on the English all-words task at Senseval-3 (Snyder and Palmer, 2004). This task required WSD systems to provide a sense choice for 2,081 content words in a set of 301 sentences from the fiction, news story, and editorial domains.

We considered the three best-ranking WSD systems – GAMBL (Decadt et al., 2004), SenseLearner (Mihalcea and Faruque, 2004), and Koc University (Yuret, 2004) – and the best unsupervised system, namely IRST-DDD (Strapparava et al., 2004). We also included SSI as it outperforms all the untrained systems (Navigli and Velardi, 2005). To evaluate the performance of the five systems on our coarse clustering, we considered a fine-grained answer to be correct if it belongs to the same cluster as that of the correct answer. Table 4 reports the performance of the systems, together with the first sense and the random baseline (in the last column we report the performance on the original fine-grained test set).

Table 4: Performance of WSD systems at Senseval-3 on coarse-grained sense inventories.

System	Prec.	Rec.	F1	$F1_{fine}$
Gambl	0.779	0.779	0.779	0.652
SenseLearner	0.769	0.769	0.769	0.646
KOC Univ.	0.768	0.768	0.768	0.641
SSI	0.758	0.758	0.758	0.612
IRST-DDD	0.721	0.719	0.720	0.583
FS baseline	0.769	0.769	0.769	0.624
Random BL	0.497	0.497	0.497	0.340

The best system, Gambl, obtains almost 78% precision and recall, an interesting figure compared to 65% performance in the fine-grained WSD task. An interesting aspect is that the ranking across systems was maintained when moving from a fine-grained to a coarse-grained sense inventory, although two systems (SSI and IRST-DDD) show the best improvement.

In order to show that the general improvement is the result of an appropriate clustering, we assessed the performance of Gambl by averaging its results when using 100 randomly-generated different clusterings. We excluded monosemous clusters from the test set (i.e. words with all the senses mapped to the same ODE entry), so as to clarify the real impact of properly grouped clusters. As a result, the random setting obtained 64.56% average accuracy, while the performance when adopting our automatic clustering was 70.84% (1,025/1,447 items).

To make it clear that the performance improvement is not only due to polysemy reduction, we considered a subset of the Senseval-3 test set including only the incorrect answers given by the fine-grained version of Gambl (623 items). In other words, on this data set Gambl performs with 0% accuracy. We compared the performance of

Table 5: Performance of SSI on coarse inventories (SSI* uses a coarse-grained knowledge base).

System	Prec.	Recall	F1
SSI + baseline	0.758	0.758	0.758
SSI	0.717	0.576	0.639
SSI*	0.748	0.674	0.709

Gambl when adopting our automatic clustering with the accuracy of the random baseline. The results were respectively 34% and 15.32% accuracy.

These experiments prove that the performance in Table 4 is not due to chance, but to an effective way of clustering word senses. Furthermore, the systems in the Table are not taking advantage of the information given by the clustering (trained systems could be retrained on the coarse clustering). To assess this aspect, we performed a further experiment. We modified the sense inventory of the SSI lexical knowledge base by adopting the coarse inventory acquired automatically. To this end, we merged the semantic interconnections belonging to the same cluster. We also disabled the first sense baseline heuristic, that most of the systems use as a back-off when they have no information about the word at hand. We call this new setting SSI* (as opposed to SSI used in Table 4).

In Table 5 we report the results. The algorithm obtains an improvement of 9.8% recall and 3.1% precision (both statistically significant, $p < 0.05$). The increase in recall is mostly due to the fact that different senses belonging to the same cluster now contribute together to the choice of that cluster (rather than individually to the choice of a fine-grained sense).

5 Related Work

Dolan (1994) describes a method for clustering word senses with the use of information provided in the electronic version of LDOCE (textual definitions, semantic relations, domain labels, etc.). Unfortunately, the approach is not described in detail and no evaluation is provided.

Most of the approaches in the literature make use of the WordNet structure to cluster its senses. Peters et al. (1998) exploit specific patterns in the WordNet hierarchy (e.g. sisters, autohyponymy, twins, etc.) to group word senses. They study semantic regularities or generalizations obtained and analyze the effect of clustering on the compatibility of language-specific wordnets. Mihalcea and Moldovan (2001) study the structure of WordNet for the identification of sense regularities: to this end, they provide a set of semantic and probabilistic rules. An evaluation of the heuristics provided leads to a polysemy reduction of 39% and an error rate of 5.6%. A different principle for clustering WordNet senses, based on the Minimum Description Length, is described by Tomuro (2001). The clustering is evaluated against WordNet cousins and used for the study of inter-annotator disagreement. Another approach exploits the (dis)agreements of human annotators to derive coarse-grained sense clusters (Chklovski and Mihalcea, 2003), where sense similarity is computed from confusion matrices.

Agirre and Lopez (2003) analyze a set of methods to cluster WordNet senses based on the use of confusion matrices from the results of WSD systems, translation equivalences, and topic signatures (word co-occurrences extracted from the web). They assess the acquired clusterings against 20 words from the Senseval-2 sense groupings.

Finally, McCarthy (2006) proposes the use of ranked lists, based on distributionally nearest neighbours, to relate word senses. This softer notion of sense relatedness allows to adopt the most appropriate granularity for a specific application.

Compared to our approach, most of these methods do not evaluate the clustering produced with respect to a gold-standard clustering. Indeed, such an evaluation would be difficult and time-consuming without a coarse sense inventory like that of ODE. A limited assessment of coarse WSD is performed by Fellbaum et al. (2001), who obtain a large improvement in the accuracy of a maximum-entropy system on clustered verbs.

6 Conclusions

In this paper, we presented a study on the construction of a coarse sense inventory for the WordNet lexicon and its effects on unrestricted WSD.

A key feature in our approach is the use of a well-established dictionary encoding sense hierarchies. As remarked in Section 2.2, the method can employ any dictionary with a sufficiently structured inventory of senses, and can thus be applied to reduce the granularity of, e.g., wordnets of other languages. One could argue that the adoption of the ODE as a sense inventory for WSD would be a better solution. While we are not against this possibility, there are problems that cannot be solved at present: the ODE does not encode semantic re-

lations and is not freely available. Also, most of the present research and standard data sets focus on WordNet.

The fine granularity of the WordNet sense inventory is unsuitable for most applications, thus constituting an obstacle that must be overcome. We believe that the research topic analyzed in this paper is a first step towards making WSD a feasible task and enabling language-aware applications, like information retrieval, question answering, machine translation, etc. In a future work, we plan to investigate the contribution of coarse disambiguation to such real-world applications. To this end, we aim to set up an Open Mind-like experiment for the validation of the entire mapping from WordNet to ODE, so that only a minimal error rate would affect the experiments to come.

Finally, the method presented here could be useful for lexicographers in the comparison of the quality of dictionaries, and in the detection of missing word senses.

Acknowledgments

This work is partially funded by the Interop NoE (508011), 6^{th} European Union FP. We wish to thank Paola Velardi, Mirella Lapata and Samuel Brody for their useful comments.

References

Eneko Agirre and Oier Lopez. 2003. Clustering wordnet word senses. In *Proc. of Conf. on Recent Advances on Natural Language (RANLP)*. Borovets, Bulgary.

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of* 3^{rd} *Conference on Language Resources and Evaluation*. Las Palmas, Gran Canaria.

Tim Chklovski and Rada Mihalcea. 2002. Building a sense tagged corpus with open mind word expert. In *Proc. of ACL 2002 Workshop on WSD: Recent Successes and Future Directions*. Philadelphia, PA.

Tim Chklovski and Rada Mihalcea. 2003. Exploiting agreement and disagreement of human annotators for word sense disambiguation. In *Proc. of Recent Advances In NLP (RANLP 2003)*. Borovetz, Bulgaria.

Bart Decadt, Véronique Hoste, Walter Daelemans, and Antal van den Bosch. 2004. Gambl, genetic algorithm optimization of memory-based wsd. In *Proc. of ACL/SIGLEX Senseval-3*. Barcelona, Spain.

William B. Dolan. 1994. Word sense ambiguation: Clustering related senses. In *Proc. of 15th Conference on Computational Linguistics (COLING)*. Morristown, N.J.

Philip Edmonds and Adam Kilgariff. 1998. Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering*, 8(4).

Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs, and Susanne Wolf. 2001. Manual and automatic semantic annotation with wordnet. In *Proc. of NAACL Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA.

Christiane Fellbaum, editor. 1998. *WordNet: an Electronic Lexical Database*. MIT Press.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In *Proc. of* 5^{th} *Conf. on Systems Documentation*. ACM Press.

Bernardo Magnini and Gabriela Cavaglià. 2000. Integrating subject field codes into wordnet. In *Proc. of the* 2^{nd} *Conference on Language Resources and Evaluation (LREC)*. Athens, Greece.

Diana McCarthy. 2006. Relating wordnet senses for word sense disambiguation. In *Proc. of ACL Workshop on Making Sense of Sense*. Trento, Italy.

Rada Mihalcea and Ehsanul Faruque. 2004. Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proc. of ACL/SIGLEX Senseval-3*. Barcelona, Spain.

Rada Mihalcea and Dan Moldovan. 2001. Automatic generation of a coarse grained wordnet. In *Proc. of NAACL Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1).

Roberto Navigli and Paola Velardi. 2004. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, 30(2).

Roberto Navigli and Paola Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7).

Hwee T. Ng, Chung Y. Lim, and Shou K. Foo. 1999. A case study on the inter-annotator agreement for word sense disambiguation. In *Proc. of ACL Workshop: Standardizing Lexical Resources*. College Park, Maryland.

Wim Peters, Ivonne Peters, and Piek Vossen. 1998. Automatic sense clustering in eurowordnet. In *Proc. of the* 1^{st} *Conference on Language Resources and Evaluation (LREC)*. Granada, Spain.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proc. of ACL 2004 SENSEVAL-3 Workshop*. Barcelona, Spain.

Catherine Soanes and Angus Stevenson, editors. 2003. *Oxford Dictionary of English*. Oxford University Press.

Christopher Stokoe. 2005. Differentiating homonymy and polysemy in information retrieval. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*. Vancouver, Canada.

Carlo Strapparava, Alfio Gliozzo, and Claudio Giuliano. 2004. Pattern abstraction and term similarity for word sense disambiguation. In *Proc. of ACL/SIGLEX Senseval-3*. Barcelona, Spain.

Noriko Tomuro. 2001. Tree-cut and a lexicon based on systematic polysemy. In *Proc. of the Meeting of the NAACL*. Pittsburgh, USA.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word sense disambiguation vs. statistical machine translation. In *Proc. of Conference on Empirical Methods in Natural Language Processing*. Vancouver, Canada.

Deniz Yuret. 2004. Some experiments with a naive bayes wsd system. In *Proc. of ACL/SIGLEX Senseval-3*. Barcelona, Spain.

Ying Zhao and George Karypis. 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3).