Natural Language Engineering **14** (4): 547–573. © 2008 Cambridge University Press
doi:10.1017/S1351324908004749 Printed in the United Kingdom

A structural approach to the automatic adjudication of word sense disagreements

ROBERTO NAVIGLI
Dipartimento di Informatica, University of Rome "La Sapienza", 00198 Rome, Italy
e-mail: navigli@di.uniroma1.it

(*Received 28 October 2006; Revised 2 December 2007; first published online 24 April 2008*)

Abstract

The semantic annotation of texts with senses from a computational lexicon is a complex and often subjective task. As a matter of fact, the fine granularity of the WordNet sense inventory [Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database* MIT Press], a *de facto* standard within the research community, is one of the main causes of a low inter-tagger agreement ranging between 70% and 80% and the disappointing performance of automated fine-grained disambiguation systems (around 65% state of the art in the Senseval-3 English all-words task). In order to improve the performance of both manual and automated sense taggers, either we change the sense inventory (e.g. adopting a new dictionary or clustering WordNet senses) or we aim at resolving the disagreements between annotators by dealing with the fineness of sense distinctions. The former approach is not viable in the short term, as wide-coverage resources are not publicly available and no large-scale reliable clustering of WordNet senses has been released to date. The latter approach requires the ability to distinguish between subtle or misleading sense distinctions. In this paper, we propose the use of structural semantic interconnections – a specific kind of lexical chains – for the adjudication of disagreed sense assignments to words in context. The approach relies on the exploitation of the lexicon structure as a support to smooth possible divergencies between sense annotators and foster coherent choices. We perform a twofold experimental evaluation of the approach applied to manual annotations from the SemCor corpus, and automatic annotations from the Senseval-3 English all-words competition. Both sets of experiments and results are entirely novel: structural adjudication allows to improve the state-of-the-art performance in all-words disambiguation by 3.3 points (achieving a 68.5% F1-score) and attains figures around 80% precision and 60% recall in the adjudication of disagreements from human annotators.

1 Introduction

Sense annotation is the task of making explicit the intended meaning of words in context. For each content word of interest, an annotator – either manual or automatic – selects an appropriate sense from a computational lexicon. This is a task where both machines and humans find it difficult to reach an agreement.

Divergent choices can be made by human annotators based on their different background, way of thinking, inherent subjectivity of the task, and especially due

to the possibly fine granularity of sense discretization. When it comes to automatic annotation, i.e. word sense disambiguation (WSD), the problem seems to get even worse: the disagreement between systems can concern completely unrelated senses, and it is more likely that gross mistakes are made.

In recent studies it has been reported that the inter-annotator agreement, i.e. the percentage of sense assignments on which the annotators agree, is between 70% and 80% (Fellbaum, Grabowski, and Landes 1998; Edmonds and Kilgarriff 2002; Snyder and Palmer 2004) on unrestricted texts when the WordNet dictionary (Fellbaum 1998) is adopted. The lack of agreement is even amplified when sense tags are collected through acquisition interfaces, due to the unknown source of the contributions of possibly unskilled volunteers (Chklovski and Mihalcea 2003).

The fine-grained nature of the WordNet sense inventory is certainly one of the major obstacles to agreed sense annotation. Unfortunately, most of the research in the Natural Language Processing community is conducted on this resource, as no other large-scale computational lexicon is freely available.

Especially when there is no clear preference toward a certain word sense, the final choice made by a judge can be subjective, if not arbitrary. This is a case where analyzing the intrinsic structure of the reference lexicon is essential for producing a consistent decision. A judge is indeed expected to review a number of related dictionary entries in order to adjudicate a sense coherently. This work can be tedious, time-consuming, and often incomplete, due to the complex structure of the resource, resulting in possibly inconsistent choices.

In this paper, we present and evaluate a method for the automatic adjudication of disagreed word senses. The approach relies on the employment of the lexicon structure as an aid to make coherent judgments. Firstly, we formalize the adjudication task (Section 2), and we introduce lexical chains and structural semantic interconnections (Section 3). Then, we illustrate our method for the adjudication of disagreed word senses (Section 4), and we evaluate the approach applied to both manual and automatic annotations (Section 5). Related work is discussed in Section 6. In Section 7 we conclude with some final remarks.

2 The adjudication task

2.1 Definition and motivation

The adjudication task can be defined as follows: let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of annotators and let σ be a sentence that each annotator in A tagged with a sense from a reference inventory (e.g. WordNet). Given a word $w \in \sigma$, we define the set of annotations provided for w by the annotators as $S_A = \{s_1, s_2, \ldots, s_m\} \subseteq Senses(w)$, where $Senses(w)$ is the set of senses of w in the reference inventory and $m \leq n$.

If $|S_A| > 1$, i.e. if at least one annotator disagreed on which sense to associate with w, an adjudication step is required. The final judgment is typically made by an adjudicator who selects for word w a sense $s \in Senses(w)$ over the others. Notice that s is a word sense for w in the sense inventory, but is not necessarily in S_A,

Table 1. *The WordNet sense inventory of the noun smell*

Sense	Hypernym	Definition
#1	sensation#1	The sensation that results when olfactory receptors in the nose are stimulated by particular chemicals in gaseous form
#2	property#3	Any property detected by the olfactory system
#3	atmosphere#1	The general atmosphere of a place or situation and the effect that it has on people
#4	sensory system#1	The faculty of smell
#5	sensing#2	The act of perceiving the odor of something

although it is likely to be. Also note that the annotators in A can be either human or automatic, depending upon the purpose of the exercise:

- **Manual annotation** is usually performed in the creation of experimental data sets (e.g. SemCor, the Senseval data sets, etc.): the possible divergencies of opinion between human annotators need to be smoothed away so as to produce high-quality data sets and guarantee that the experiments will provide meaningful results;
- **Automatic annotation** is gaining more and more interest with the advent of the so-called Semantic Web (Berners-Lee 1999), as users are not willing to sense tag and semantically index their web pages. However, compared to human annotators, automated systems tend to diverge in their choices more often and at a coarser level of granularity. As a result, adjudicating these disagreements is a laborious and time-consuming task.

As an example, suppose that human annotators manually tagged the following sentence (we subscript the sense annotations beside each word):[1]

(a) She loves$_{\#1}$ the smell$_{\#1,2}$ of basil$_{\#1}$ leaves$_{\#1}$

The uncertainty on the appropriate choice for the noun *smell* is reflected by the ambiguity of its WordNet sense inventory, reported in Table 1. This is a case where even a good judge might not be able to choose between the first two senses of the word, even knowing the hypernyms of both (as reported in the table). However, a closer look at the lexicon structure would reveal that the appropriate sense of *smell* is #1, as shown in Figure 1. In fact, the second sense of *smell* expresses one of the abstract properties typical of human beings (vision, audition, touch, taste, smell). This information (the taxonomical structure, but also other important textual, lexical, and semantic information) is not immediately available to the annotator. As a result, inconsistent choices can be made.

[1] 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. For the sake of clarity, we omit the part of speech where it can be deduced from the context.

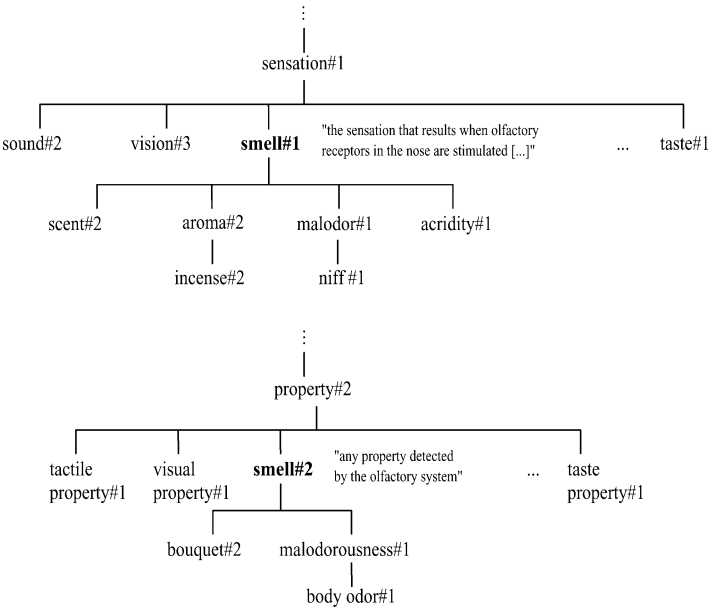

Fig. 1. An excerpt of the WordNet taxonomy including *smell#1* and *smell#2*.

2.2 Automatizing the adjudication task

Adjudication is usually performed by a human judge. In this paper, we propose the automatization of the adjudication task. A way to look at automatic adjudication is as a special case of WSD. Given a sentence σ, we apply a WSD algorithm to σ by taking advantage of the following two conditions:

(1) The set of agreed word senses can be used as a fixed semantic context to help the disambiguation of senses with disagreement;
(2) For each disagreed word $w \in \sigma$, the appropriate meaning of w is selected from the subset of senses chosen by the annotators.

This setting seems more favorable compared to the general disambiguation task, in which we discard senses that are likely to provide noisy information (condition (1)), and senses that were not chosen by any of the annotators (condition (2)). However, notice that WSD *per se* is a very difficult task and that the adjudication task often concerns hard cases like extremely subtle sense distinctions and multiple interpretations, rather than simple cases of homonymy (we further discuss this phenomenon in Section 5.1.2).

For example, the automatic adjudication of the disagreed word *smell* in sentence (a) can be performed by fixing the word senses *love#v#*1, *basil#n#*1, and *leaf#n#*1 (as of condition (1)), and by applying a WSD algorithm to the disagreed noun *smell* on the restricted sense inventory {*smell#n#*1, *smell#n#*2}.

Most WSD algorithms cannot take full advantage of the above-mentioned disambiguation setting. Supervised approaches rely on an adequate amount of training data (a heavy assumption indeed, because of the scarce availability of annotated data, especially in case of automatic annotation, and the intrinsic

impossibility of guaranteeing the quality of the training data with the very same approach). Moreover, these approaches are usually trained on lexico-syntactic features and find it difficult to exploit explicit semantic hints (condition (1)) to improve their disambiguation quality. This is not only due to the frequent lack of contextual semantic clues in training data sets, but also to the difficulty of taking into account a variable number of clues.

Ensembles are probably the class of algorithms which best approximates our conditions for adjudication. However, these algorithms again do not benefit from the fixed semantic context provided by agreed senses and often need to be trained. We provide further discussion on these approaches in Section 6.

These remarks lead us to focus on knowledge-based WSD algorithms, which can satisfy both conditions (1) and (2). A further desideratum is to select an algorithm which is able to make choices which are coherent with the lexicon structure, as the adjudication must be justifiable and, if possible, further analyzed by human judges. We present hereafter a structural approach to the automatic adjudication of disagreed sense choices that fits the above-mentioned requirements.

3 Lexical chains and semantic interconnections

Lexical chains (Morris and Hirst 1991) are sequences of words $w_1, \ldots, w_n$ in a text that represent the same topic, i.e. such that w_i is related to w_{i+1} by a lexico-semantic relation (e.g. hypernymy, meronymy, etc.).

Lexical chains have been applied to the analysis of discourse cohesion (Morris and Hirst 1991), text summarization (Barzilay and Elhadad 1997), the correction of malapropisms (Hirst and St-Onge 1998), WSD (Galley and McKeown 2003), etc. This idea was developed in further approaches to WSD based on lexico-semantic heuristics (Rigau, Atserias, and Agirre 1997; Harabagiu, Miller, and Moldovan 1999) and link analysis (Mihalcea, Tarau, and Figa 2004; Véronis 2004; Agirre, Martínez, de Lacalle, and Soroa 2006).

Recently, a knowledge-intensive, untrained algorithm for WSD, called *Structural Semantic Interconnections*[2] (SSI) (Navigli and Velardi 2005), has been shown to provide interesting insights into the choice of word senses by producing structural justifications in terms of semantic graphs.

SSI 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 (Lea 2002), the Longman Language Activator (Longman 2003), and collocation web sites]. Each collocation is mapped to the WordNet sense inventory in a semiautomatic manner and transformed into a *relatedness* edge [for further details, the interested reader can refer to Navigli (2005)]. Notice that, at present,

[2] SSI is available online from `http://lcl.uniroma1.it/ssi`.

Table 2. *The full context-free grammar for the recognition of semantic interconnections*

		Pattern rules	
S	$\rightarrow$	$S_1 \mid S_2$	(start rule)
S_1	$\rightarrow$	$PHRHRHP$	(relatedness, hypernymy, meronymy)
S_2	$\rightarrow$	$NSASN$	(additional relations)
		Basic non terminals	
H	$\rightarrow$	$e_{kind-of} H \mid \epsilon$	(hypernymy)
H	$\rightarrow$	$e_{part-of} H \mid e_{has-part} H$	(meronymy)
R	$\rightarrow$	$e_{related-to} \mid \epsilon$	(relatedness)
		Additional relations	
P	$\rightarrow$	$E \mid N \mid S \mid A$	(additional relations, inter-part-of-speech, etc.)
E	$\rightarrow$	$e_{entails} \mid e_{cause} \mid \epsilon$	(entailment, cause)
N	$\rightarrow$	$e_{pertains-to} \mid e_{attribute} \mid \epsilon$	(pertainymy, attribute)
S	$\rightarrow$	$e_{similar-to} \mid e_{see-also} \mid \epsilon$	(similarity, see-also)
A	$\rightarrow$	$e_{antonym} \mid \epsilon$	(antonymy)

the lexical knowledge base does not include any information from the SemCor semantically annotated corpus (Miller, Leacock, Tengi and Bunker 1993).[3]

Given a word context $W = \{w_1, \ldots, w_k\}$, SSI builds a graph $G = (V, E)$ such that $V = \bigcup_{i=1}^{k} Senses(w_i)$ (i.e. V includes a vertex for each sense of a word in W) and $(s, s') \in E$ if there is at least one semantic interconnection between senses 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. The full context-free grammar encoding semantic interconnection patterns for the WordNet lexicon is reported in Table 2. For further details, the reader can refer to (Navigli and Velardi 2005).

The rules in the table are divided into three groups: pattern rules, basic nonterminals (identifying basic sequences of hypernymy, meronymy, and relatedness), and additional relations (most of which connect different parts of speech, like nominalization, pertainymy, etc.). The grammar is general enough to be applied to any sufficiently structured computational lexicon, where only the last group should be replaced by the set of lexicon-specific nonterminals.

The purpose of the grammar is twofold: first, to prune out a large number of unwanted lexical chains (e.g. *universe#1* $\xrightarrow{kind-of}$ *natural object#1* $\xrightarrow{kind-of}$ *object#1* $\xrightarrow{has-kind}$ *commodity#1* $\xrightarrow{has-kind}$ *merchandise#1*); second, to ensure that the chains contain a maximum number of edges of a certain kind (e.g. the grammar does not allow any sequence of three relatedness edges, like in *job#1* $\xrightarrow{related-to}$ *money#1* $\xrightarrow{related-to}$ *coin#1* $\xrightarrow{related-to}$ *metal#1*). Both aspects aim at avoiding undesired shifts of meaning.

[3] As a result, experiments performed on SemCor in later sections are not biased by the use of information from the corpus itself.

Table 3. *Good and bad examples of semantic interconnections*

Semantic interconnection	Good?
$eat\#v\#1 \xrightarrow{cause} feed\#v\#2 \xrightarrow{related} food\#n\#2$	✓
$drive\#v\#1 \xrightarrow{related} vehicle\#n\#1 \xrightarrow{related} fender\#n\#1 \xrightarrow{part-of} car\#n\#1$	✓
$cup\#n\#2 \xrightarrow{related} milk\#n\#1 \xrightarrow{related} beverage\#n\#1 \xrightarrow{has-kind} coffee\#n\#1$	✓
$shivery\#a\#1 \xrightarrow{similar-to} cold\#a\#1 \xrightarrow{attribute} temperature\#n\#1$	✓
$computer\#1 \xrightarrow{related} user\#1 \xrightarrow{kind-of} consumer\#1 \xrightarrow{has-kind} drinker\#2$	×

Even though the grammar produces an infinite number of edge sequences, we limit the recognition to strings[4] of length ≤ 5. While it is true that this reduces the expressivity of the grammar to a finite state automaton (or, equivalently, a regular grammar), yet we want to maintain a general, clear, and compact formalism to express edge sequences.

We report good and bad examples of semantic interconnections in Table 3. In the second column, we mark each interconnection in the table with a check mark (✓) if the interconnection connects two concepts which are really semantically related (we mark it with ×, otherwise).

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

$$Score_{\mathrm{SSI}}(s, \mathscr{C}) = \frac{\sum_{s' \in \mathscr{C} \setminus Senses(w)} \sum_{i \in IC(s,s')} \frac{1}{length(i)}}{\sum_{s' \in \mathscr{C} \setminus Senses(w)} |IC(s, s')|}$$

where $IC(s, s')$ is the set of interconnections between senses s and s'. The contribution of a single interconnection i is given by the reciprocal of its length ($1/length(i)$), 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 $\hat{s}_w$ is chosen, the other senses of w are removed from the semantic context $\mathscr{C}$ (i.e. $\mathscr{C} \leftarrow \mathscr{C} \setminus Senses(w) \cup \{\hat{s}_w\}$), and the word w, disambiguated as a result of the current iteration, is removed from W. During the next iteration, $\hat{s}_w$ will be used as an additional semantic context for the remaining words in W, i.e. for the words yet to be disambiguated. The algorithm terminates when either $\mathscr{C} = \emptyset$ or there is no sense such that its score exceeds a fixed threshold. The threshold was experimentally set to 0.2.

[4] This is done for two reasons: a computational aspect, and the idea that longer sequences are more likely to lead to an undesired semantic shift. Experiments for tuning this parameter, as well as the termination threshold introduced below, were performed based on the performance of SSI on an in-house manually annotated data set.

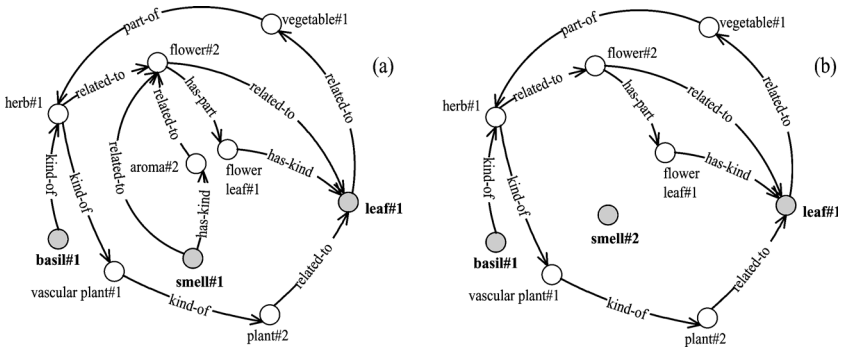

Fig. 2. (a) Some semantic interconnections supporting the choice of sense #1 of *smell* in sentence (a). (b) The choice of *smell#2* is not supported by any semantic interconnection.

4 Adjudication with structural semantic interconnections

In this section, we illustrate the use of structural semantic interconnections for the automatic adjudication of the most appropriate word sense in case of disagreement.

Given a sentence σ and a set of words with disagreement $W \subseteq \sigma$, according to Section 2.2, we apply SSI to W by taking into account for disambiguation only the senses selected by the annotators, and using as a fixed context the agreed senses chosen by the annotators for the words in $\sigma \setminus W$.

Recall the annotated sentence from section 2.1:

(a) She $\text{loves}_{\#1}$ the $\text{smell}_{\#1,2}$ of $\text{basil}_{\#1}$ $\text{leaves}_{\#1}$

This sentence is a real case from an annotation experiment we conducted in a previous work. We initialize the word context $W = \{smell\#n\}$ and the semantic context $\mathscr{C} = \{\text{love\#v\#1, basil\#n\#1, leaf\#n\#1}\} \cup \{\text{smell\#n\#1, smell\#n\#2}\}$ (the first set is the fixed semantic context of agreed senses, while the second set contains the disagreed word senses). Only the first two senses of *smell* (i.e. the disagreed senses) are taken into account for the calculation of the $Score_{\text{SSI}}$, leading to:

$$Score_{\text{SSI}}(smell\#n\#1, \mathscr{C}) = 0.80$$

$$Score_{\text{SSI}}(smell\#n\#2, \mathscr{C}) = 0$$

The choice of the first sense of *smell* as a solution to this disagreement is structurally supported by a number of semantic interconnections according to the grammar in Table 2. Figure 2(a) shows some interconnections identified by the algorithm. In contrast, sense 2 is not related to other senses in $\mathscr{C}$ through any semantic interconnection, as illustrated in Figure 2(b).

As a second example, consider the WordNet definition of *motorcycle*:

(b) Motorcycle: a motor vehicle with two wheels and a strong frame

In the Senseval-3 Gloss WSD task (Litkowski 2004), the human annotators assigned the first sense to the word *frame* (*a structure supporting or containing something*), unintentionally neglecting that the dictionary encodes a specific sense of *frame* concerning the structure of objects (e.g. vehicles, buildings). In fact, according

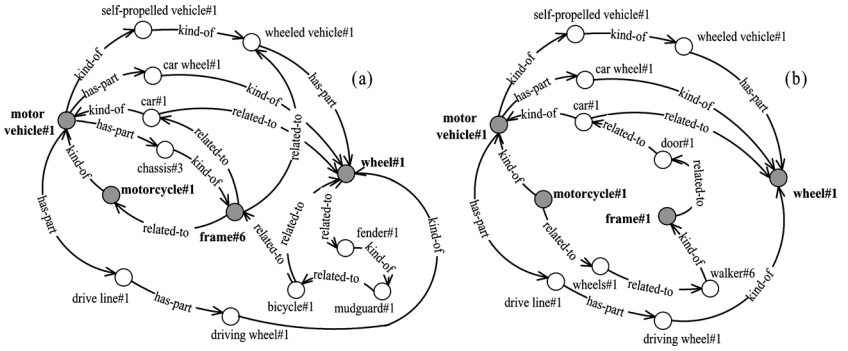

Fig. 3. (a) Some semantic interconnections supporting the choice of sense #6 of *frame* in sentence (b). (b) The choice of *frame#1* has a weaker structural support within the SSI lexical knowledge base.

to WordNet, a *chassis*#3 is a kind of *frame*#6 (*the internal supporting structure that gives an artifact its shape*), and is also part of a *motor vehicle*#1. Figures 3(a) and (b) illustrate the graphs resulting from the sense choices of *frame* #6 and #1, respectively.

Semantic interconnections reflect the fine granularity of the inventory, as they are expressions of the lexical knowledge base from which they are extracted. In fact, the choice of *frame*#1 still produces relevant semantic interconnections, as illustrated in Figure 3(b), but the overall ranking of this sense selection, i.e. the degree of overall connectivity of the resulting graph, is smaller than that obtained for *frame*#6:

$$Score_{\mathrm{SSI}}(frame\#n\#6, \mathscr{C}) = 0.65$$

$$Score_{\mathrm{SSI}}(frame\#n\#1, \mathscr{C}) = 0.53$$

These two real-world cases make it evident that semantic interconnections can point at inconsistent, though acceptable, choices made by human annotators due, among others, to the fine granularity of the sense inventory and to regular polysemy, i.e. the recurring and predictable sense alternations certain classes of words are subject to.

We recognize that subtle distinctions, like those encoded in WordNet, are rarely useful in any NLP application, but as a matter of fact WordNet is at the moment the *de facto* standard within the research community, as no other computational lexicon of that size and complexity is freely available.

5 Evaluation

In this section, we present an evaluation of the effectiveness of semantic interconnections applied to the adjudication of disagreed word senses. We assessed the method for both manual (Section 5.1) and automatic annotations (Section 5.2).

5.1 Evaluating the adjudication of manual annotations

As mentioned above, the adjudication of disagreements resulting from manual annotations is a critical task when we want to guarantee the high quality of a data set. To assess the quality of the adjudications made by SSI on manual sense annotations, we performed two different experiments on SemCor (Miller *et al.* 1993), a corpus of more than 200,000 content words manually tagged with WordNet word senses:

- A large-scale simulation of disagreements between two annotators (Section 5.1.1);
- An experiment on a smaller scale concerning the adjudication of real disagreements between annotators of a portion of the SemCor corpus (Section 5.1.2).

5.1.1 Adjudicating simulated disagreements

As a first experiment, we simulated *in vitro* a disagreement between two annotators in which an annotator provides an appropriate sense and the other selects a different sense at various semantic levels. To this end, we needed a way to distinguish meanings at different levels of granularity. For each word of interest, we used the *Oxford Dictionary of English* (ODE) (Soanes and Stevenson 2003) to manually produce a hierarchical version of the WordNet sense inventory of that word.

The ODE[5] provides a multilevel 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 *microdistinctions* (e.g. a division of humankind versus a group of people with a common ancestor). Table 4 shows the lexical entries of the noun *race* (we represent an ODE sense of a word w as $w\#p\#h.i$, where p is its part of speech and i denotes the i-th polysemous entry of the h-th homonym of w).

For each word of interest, the manual mapping of WordNet senses to ODE polysemous entries induces a hierarchical structure on the former. For example, consider the sense inventories provided by the two dictionaries for the noun *race* (Tables 4 and 5). As a result of the mapping, a semantic correlation between the WordNet senses is identified at different levels of granularity, as shown in Figure 4.

We randomly selected a set W of 207 polysemous words from WordNet (64 nouns, 82 verbs, 61 adjectives) resulting in 1,488 different senses overall (7.18 senses per word on average) that we manually mapped to the appropriate ODE senses.[6] Notice that mapping fine-grained to coarse-grained word senses is much easier than any semantic annotation or one-to-one mapping task. This intuition is also substantiated by a quantitative assessment: 548 WordNet senses of 60 words were mapped to ODE entries by two annotators, with a pairwise agreement of 92.7% (κ agreement: 0.854).

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

[6] In the experiments, we neglected adverbs as very few interconnections can be found for them.

Table 4. *The sense inventory of race#n in the ODE*[*]

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).

[*]definitions are abridged, bullets (•) indicate a subsense in the ODE, arrows (→) indicate hypernymy.

Table 5. *The sense inventory of race#n in WordNet*[*]

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).

[*]arrows (→) indicate hypernymy.

This mapping technique was also successfully used in the organization of the Semeval-2007 coarse-grained English all-words WSD task (Navigli, Litkowski, and Hargraves 2007). An expert lexicographer estimated that an amount of time between 20 and 30 s is enough for mapping a fine-grained sense (e.g. from WordNet) to a coarser sense (e.g. in ODE) with the aid of an appropriate interface. Accordingly, the creation of our mapping took an overall amount of 11 man-hours.

Next, we selected all those sentences in SemCor which included at least one occurrence of words from W (overall, 183 of the 207 words in W occur at least once and are sense-tagged in SemCor). Formally, we considered all the sentences $\sigma = w_1 w_2 \dots w_n$ annotated in SemCor with the senses $s_{w_1} s_{w_2} \dots s_{w_n}$ ($s_{w_j} \in Senses(w_j)$,

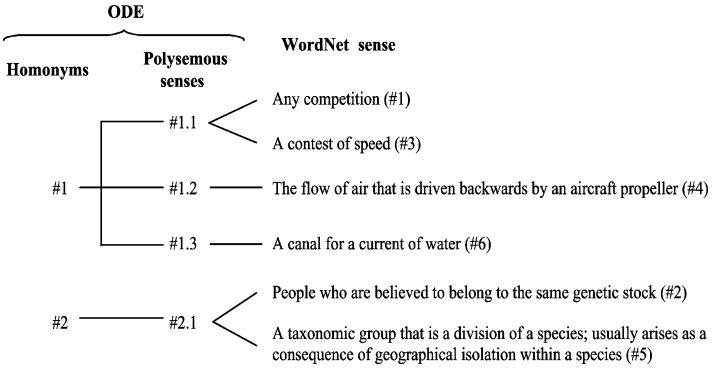

Fig. 4. A hierarchical version of the WordNet sense inventory for noun *race*.

Table 6. *Size and composition of the four disagreement test sets*

Granularity	Nouns	Verbs	Adjectives	Total	# Sentences
Microdistinctions	1,885	2,829	726	5,440	4,542
Polysemy	1,453	2,701	532	4,686	3,996
Homonymy	437	183	49	669	633
Random disagreements	1,933	2,860	774	5,567	4,633

$j \in \{1, 2, \ldots, n\}$), such that σ contains at least one word $w_i \in W$. Then, for each sentence we simulated a disagreement on the word w_i by randomly selecting a different sense $\overline{s}_{w_i}$ for the word w_i at three different levels:

(1) **microdistinctions**: $\overline{s}_{w_i}$ is in the same lexical entry as s_{w_i} (e.g. *race*#n#1 and *race*#n#3, both mapped to the ODE entry *race*#n#1.1) – we expect that most of the human disagreements fall into this category (as also confirmed by the experiments in Section 5.1.2);
(2) **polysemy**: $\overline{s}_{w_i}$ is in a different polysemous entry from that of s_{w_i} (e.g. *race*#n#1 and *race*#n#4, mapped respectively to the ODE polysemous entries *race*#n#1.1 and *race*#n#1.2);
(3) **homonymy**: $\overline{s}_{w_i}$ is a homonym of s_{w_i} (e.g. *race*#n#1 and *race*#n#2, mapped respectively to the ODE homonyms *race*#n#1.1 and *race*#n#2.1).

We further experimented on a random choice of the disagreed sense:

(4) **random** disagreements: we built a fourth data set in which sense $\overline{s}_{w_i}$ of word w_i is chosen randomly from $Senses(w_i) \setminus \{s_{w_i}\}$.

In Table 6 we report the size of the four test sets, i.e. the number of disagreements by part of speech (and overall) and the total number of sentences involved, ranging from 4,633 sentences for random disagreements (i.e. the full set of sentences which contain at least one word from our initial set W of 207 words) to 633 for homonymy (i.e. the set of sentences which include at least a homonym). This difference is due to the low number of homonyms compared to polysemous entries. Notice that the

Table 7. *Performance of automatic adjudication at different levels of disagreement (sentences from SemCor)*

Granularity	Precision	Recall	F1-score	# Instances
Microdistinctions	76.46	57.63	65.72	5,440
Polysemy	82.94	60.71	70.11	4,686
Homonymy	86.10	75.93	80.70	669
Random disagreements	80.78	60.54	69.21	5,567

number of disagreements per sentence is a function of the initial set of words W whose senses were clustered according to ODE. As a consequence, the resulting ratio (ranging from 1.05 to 1.2 disagreements per sentence depending upon the class of granularity) can differ significantly from that of a real data set of human disagreements. This aspect is further discussed in Section 5.1.2. Finally, the average number of context (i.e. agreed) words per sentence ranges between ten and twelve words depending on the data set.

We applied SSI to the annotated sentences (as discussed in Sections 2.2 and 4) and evaluated the performance of the approach in suggesting the appropriate choice for the words with disagreement. We assessed *precision* (the number of correct suggestions over the overall number of suggestions from SSI) and *recall* (the number of correct suggestions over the total number of words to be adjudicated), and we calculated the F1-score, a harmonic mean of the two measures ($\frac{2 \cdot p \cdot r}{p+r}$). The results are reported in Table 7.

As expected, the accuracy of automatic adjudication increases as the "semantic distance" of the disagreed senses grows. In fact, while we get an F1-score of 65.72% in the adjudication of microdistinctions, this figure rises to 80.7% when we deal with homonyms. This confirms the intuitive idea that closer senses are inherently difficult to disambiguate.

Moreover, it is interesting to note that there is a 6.5% increase in precision from microdistinctions to polysemy and a 3.1% increase from polysemy to homonymy. As expected, we observe the largest increase in recall from polysemy to homonymy (+15.22%). Completely distinct senses of the same word convey indeed very different semantics and it is less likely to find semantic interconnections supporting the wrong sense.

We observe that the performance on randomly selected disagreements is close to that obtained for polysemous sense distinctions (indeed, the difference is not statistically significant according to a χ^2 test). We will see in Section 5.2 that random disagreements apparently simulate the average level of difficulty encountered by an automatic system.

We preferred to calculate precision and recall distinctly, rather than accuracy, as the latter does not provide any hint about the quality of the structural suggestions produced by SSI: a good, justifiable adjudication is better than a couple of mediocre answers.

Table 8. *Performance by part of speech of automatic adjudication of subtle disagreements (sentences from SemCor)*

Part of speech	Precision	Recall	F1-score	# Instances
Nouns	76.84	72.52	74.62	1,885
Verbs	75.84	53.27	62.58	2,829
Adjectives	78.14	35.95	49.24	726
Overall	76.46	57.63	65.72	5,440

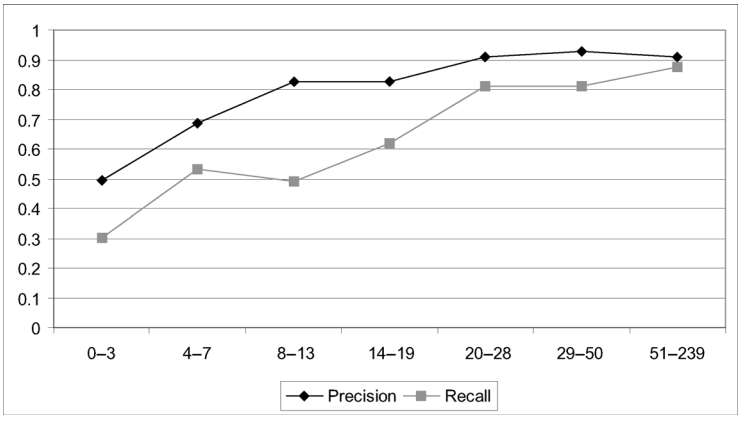

Fig. 5. The SSI performance on the microdistinction test set by degree of connectivity. Each interval includes a minimum of 300 word instances from the test set.

We studied the performance by part of speech for the most difficult case, i.e. disagreements on microdistinctions (similar trends were observed for the other classes of granularity). We report the figures in Table 8. The most interesting result can be observed on noun performance: here SSI proves to be really effective, by correctly adjudicating almost three-fourth of the disagreements. The drop in recall for verbs and, even more, for adjectives is mostly due to the lack of structural information. However, precision keeps high in both cases, which allows for high-quality adjudication and leaves the remaining uncertain cases to a more careful, manual inspection.

We also calculated the accuracy of the most frequent sense (MFS) heuristic for the four levels of granularity: 89.54% (microdistinctions), 93.85% (polysemy), 98.95% (homonymy), and 92.17% (random). The MFS baseline selects the disagreed sense which is most frequent in the SemCor annotated corpus. Consequently, it cannot be considered as a baseline, as it employs the test corpus to determine its sense choices. However, it can be used to establish an upper bound for this adjudication task.

Finally, we studied how the degree of connectivity of each word instance affects the performance of SSI on the adjudication task. Specifically, we extrapolated the trend on the microdistinction test set, that we report in Figure 5 (a similar trend was observed for the other levels of granularity). On the x-axis we have the out-degree of the correct sense of each word instance in the test set. The out-degree counts the

number of WordNet relation edges and *relatedness* edges (cf. Section 3) outgoing from the word sense at hand. Each interval in the figure includes a minimum of 300 word instances from the test set, with approximately two-thirds of the word instances distributed in the range 0–13.

The graph shows that the higher the degree of connectivity in the computational lexicon, the better is the overall performance of structural semantic interconnections (both in terms of precision and recall). When we have more than eight to thirteen relation edges that enable the connectivity from a specific sense to other senses in the lexicon, the performance grows noticeably. The trend shown in the graph corroborates the apparently trivial observation that the availability of large knowledge bases (even (semi)automatically acquired) tends to improve the disambiguation performance (Cuadros and Rigau 2006). Moreover, it shows that the improvement can be really significant, with an increase of several tenths of points in terms of both precision and recall.

5.1.2 Adjudicating real disagreements

Performing experiments on a data set of simulated disagreements has the advantage of providing results on a large scale. However, experimenting on a data set of real, human disagreements would make the results more meaningful. Moreover, it would allow us to understand the level of disagreement attained by sense taggers according to the different classes of granularity that we introduced in the previous section.

Unfortunately, only few data sets of human disagreements are available to the research community. Among these, we cite the outcome of the Open Mind Word Expert (OMWE) project (Chklovski and Mihalcea 2002). However, this annotation effort does not fit our problem, as it was created on a lexical sample basis, whereas we expect annotators to tag all words in a text. Another interesting data set of disagreements is an outcome of the MultiSemCor project (Pianta, Bentivogli, and Girardi 2002; Bentivogli, Forner, and Pianta 2004), aiming at the creation of a parallel English/Italian version of the SemCor corpus. This data set includes about a hundred disagreements; as a result, its sample size does not allow to perform a statistically significant evaluation.

Given the lack of all-words disagreement data sets, we asked two annotators to manually sense-tag a portion of the SemCor corpus with WordNet senses (we adopted version 2.0). The data set included an overall number of 2,927 words (285 sentences, 2,274 polysemous words) from three randomly chosen documents on different topics, specifically, football (document code: `br-a13`), biography (`br-e22`), and mathematics (`br-j19`).

We measured a pairwise interannotator agreement of 81.1% (i.e. the annotators – expert lexicographers – disagreed on 553 words). This figure is only slightly higher than the inter-tagger agreement ratio reported by Fellbaum *et al.* (1998) on a subset of SemCor (ranging between 72.8% and 79.9% under different conditions), corroborating the difficulty of annotating with fine-grained word senses. However, in their experiments several factors lowered the agreement: the use of several annotators (with a comparison of the agreement among nonexperts and between experts and

Table 9. *Some of the inexact sense tags in the SemCor corpus*

SemCor sentence	Correct sense	SemCor sense tag
...his Credo of words – **torrents** of powerful music.	3. an overwhelming number or amount	1. a heavy rain
...never a **slave** to its academic dialectics.	3. someone entirely dominated by some influence or person	1. a person who is owned by someone
...to **cultivate** his musical talent...	3. train to be discriminative in taste or judgment	1. foster the growth of
a young giant dwindled in stature and **fruitfulness**.	2. the intellectual fruitfulness of a creative imagination	1. the quality of something that causes or assists healthy growth
...he stretches the limits of **instrumentation** with good judgement...	3. the instruments called for in a musical score or arrangement for a band or orchestra	2. the act of providing or using the instruments needed for some implementation
Prokofieff was **guided** in a consistent direction by life of...	3. be a guiding force, as with directions or advice	2. take somebody somewhere

nonexperts), an older sense inventory (several sense entries in WordNet now report SemCor sentences as examples, which make it easier for an annotator to select the appropriate meaning), etc. We do not provide here a figure of κ agreement (Cohen 1960) as for most words we have a very low number of occurrences. As a result, the κ agreement is below 0.60 (i.e. between moderate and poor agreement) for over 90% of the words.

After an adjudication step, performed by a human arbiter, we discovered that the 553 disagreements included 38 incorrect sense choices due to inattention (i.e. unvoluntary choices) of either of the two annotators, and 22 words tagged with a different part of speech than that assigned in SemCor. Moreover, the arbiter found out that 92 of the remaining disagreements had to be resolved in contrast with the corresponding senses chosen in the SemCor data set. We can say that most of these cases are due to imprecise choices in SemCor (we report some in Table 9).

Our aim is not to criticize SemCor, which is an invaluable resource for WSD and related fields. Most of these imprecisions indeed may be due to several factors, such as the mappings between different versions of WordNet, typos, and agreements by chance of the original annotators.

In order to study the distribution of the disagreements according to the three classes of sense granularity introduced in Section 5.1.1, and similarly to what we did in that section, we mapped all disagreed senses (770 distinct senses overall) to the corresponding ODE senses.

The application of SSI to solve the 553 disagreements in our data set led to the figures in Table 10. We report the overall performance of SSI together with its performance on the three classes of sense granularity. We registered only one case

Table 10. *Performance of automatic adjudication on real human disagreements at different levels of disagreement*

Granularity	Precision	Recall	F1-score	# Instances
Microdistinctions	78.32	59.73	67.77	375
Polysemy	80.15	63.37	70.78	172
Homonymy	—	—	—	1
Not mappable	—	—	—	5
Overall	78.60	61.12	68.77	553
MFS baseline	69.98	69.98	69.98	553

Table 11. *Performance of automatic adjudication on real human disagreements by part of speech*

Part of speech	Precision	Recall	F1-score	# instances
Nouns	78.73	67.41	72.63	313
Verbs	78.74	59.88	68.03	167
Adjectives	77.14	38.57	51.43	70
Overall	78.60	61.12	68.77	553

of homonymous disagreement, due to a tagging mistake (a low figure was expected indeed). We were not able to determine the granularity of five disagreements because of missing entries in ODE.

First, we note that the overall performance (68.77% F1) gets very close to the MFS baseline (69.98% F1). As in the previous experiment, we remark that the MFS baseline employs sense frequencies from the SemCor annotated texts of which our test set is a part. Secondly, the overall performance is comprised between the results of microdistinctions and the polysemous case of our simulated experiment (cf. Table 7). This result was expected, as the disagreements concerned these two classes of granularity.

Regarding the performance on distinct parts of speech, reported in Table 11, we observe an increase in recall for verbs and adjectives compared to the figures from our simulated experiment on microdistinctions (cf. Table 8). However, all the differences in precision from Tables 8 and 11 calculated distinctly for each part of speech are not statistically significant according to χ^2 tests.[7]

We remark here that the results of this experiment on real human disagreements are not entirely comparable with those from our previous experiment on simulated disagreements. Firstly, here we are forcibly working on a smaller number of disagreed senses. Unfortunately, as we mentioned above, no medium-scale data set of disagreements from all-words annotation efforts is available, so we believe our data set

[7] For information on how to perform χ^2 tests on difference sample sizes, see, e.g., Miller and Miller (2003).

Table 12. *Distribution of real human disagreements over a subset of SemCor (285 sentences)*

Disagreements per sentence	# Sentences	# Instances
0	53	0
1	69	69
2	81	162
3	41	123
4	25	100
> 4	16	99
Total	285	553

constitutes an interesting effort in this direction. Secondly, the simulated experiment was easier than this experiment on human disagreements, as it presupposed a disagreement per sentence ranging from 1.05 to 1.2 (depending upon the class of granularity). In contrast, as shown in Table 12, we can find between one and six disagreements per sentence in our data set of real annotations, with an average of 1.94 disagreements per sentence, thus making automatic adjudication a harder task, as its difficulty depends on the quality and number of agreed senses. We did not determine the performance by number of disagreements per sentence because of the small size of each sample.

5.2 Evaluating the adjudication of automatic annotations

For assessing semantic interconnections applied to the adjudication of disagreements between automatic systems, we chose the Senseval-3 corpus for the English all-words task (Snyder and Palmer 2004). The task required WSD systems to provide a sense choice for a total of 2,041 content words in a set of 301 sentences from the fiction, news story, and editorial domains.

For our experiments, we focused on the outcome of the three best-ranking systems: GAMBL (Decadt, Hoste, Daelemans, and van den Bosch 2004), SenseLearner (Mihalcea and Faruque 2004), and Koc University (Yuret 2004). The application of SSI to the disagreement set led to the figures in Table 13, where we compare our results with:

- the **chance baseline**, calculated as the sum of the uniform probabilities of correctly solving each disagreement divided by the total number of disagreements;
- the **most frequent sense baseline**, i.e. the choice of the most frequent or predominant sense in SemCor, selected among those output by the three systems;
- a **majority voting** combination strategy, i.e. the sense output by a majority of the three systems is chosen – in case of full disagreement, a random choice is made;
- the **best-performing Senseval-3 system**, i.e. GAMBL;

Table 13. *Results on the Senseval-3 all words task*

	Precision	Recall	F1
SSI	68.7	68.3	68.5
Chance baseline	63.8	63.8	63.8
MFS baseline	61.9	61.9	61.9
Majority voting	65.2	65.2	65.2
GAMBL (best system)	65.2	65.2	65.2
Oracle (upper bound)	76.9	76.9	76.9

Table 14. *The oracle performance by kind of agreement between the three best-ranking systems at Senseval-3*

	Correct	Total	Accuracy
Three-way agreement	1,082	1,386	78.1
Two-way agreement	422	557	75.8
No agreement	66	98	67.3
Total	1,570	2,041	76.9

- the **oracle** performance: if we had an oracle adjudicating the appropriate answer for each word instance, its accuracy would be 76.9% (this percentage represents the number of words for which at least one correct answer was provided by any of the three systems). This figure constitutes an upper bound for our task, as it would be impossible for any automatic adjudication algorithm to assign a greater number of correct answers.

In Table 14 we report the oracle accuracy based on the kind of agreement (three-way agreement, two-way agreement, or total disagreement). This is given by the percentage of word occurrences for which at least a correct answer is given by any of the three systems. Clearly, the oracle is more accurate when the three systems agree (78.1%), while its accuracy decreases to 67.3% when they fully disagree, indicating an inherent condition of difficulty.

Notice that the chance and MFS heuristic, as well as the majority voting strategy, have been applied to the adjudication of disagreed senses, while the performance of the best-performing system is from the original Senseval-3 task. It could be argued that, in the majority voting strategy, in case of tie the sense of the best system (namely, GAMBL) could be selected. We tested this option and obtained a result which is comparable (65.1%) with the random backoff strategy. This result can be probably attributed to the lack of information in GAMBL for assigning the appropriate sense to these challenging instances. We also experimented with a higher number of systems, and found out that the improvement in accuracy was negligible (+0.5% with five best-ranking systems), compared to the costly requirement of a higher number of state-of-the-art systems to combine.

Table 15. *Results on the Senseval-3 all words task by part of speech*

Part of speech	Precision/recall
Nouns	74.5
Adjectives	71.4
Verbs	59.5

We could not apply other kinds of untrained combination strategies, such as rank-based or probability mixture, as the confidence of state-of-the-art systems on each word sense was not available. However, experimentally, we do not expect a big increase in performance (see, e.g. Brody, Navigli, and Lapata (2006), where the performance increase is lower than 1% with a rank-based ensemble compared to majority voting, both applied to disagreed nouns from the entire SemCor).

SSI applied to automatic adjudication performs better than the best-performing Senseval system, with a difference in precision and recall of 3.5% and 3.1%, respectively. Both differences are statistically significant (χ^2 test, $p < 0.05$).

An interesting remark is that the F1 performance of SSI is not dissimilar from that obtained in the previous section with the random selection of disagreed senses. This result is plausible given that automated systems can make very good choices as well as provide inexplicable answers (especially when they disagree).

Finally, Table 15 reports the precision of SSI on nouns, adjectives, and verbs. The good performance on nouns is not surprising, as SSI relies on the lexicon structure, which is more interconnected for nouns. In contrast, verbs suffer from the lower degree of connectivity with other parts of speech, especially because of short sentences in the test set such as "that's what the man said", "I'm just numb", etc. The good performance on adjectives is given by their low degree of polysemy.

6 Related work

It is a matter of fact that achieving a high quality for sense annotations is a very hard task (see, e.g., Hanks (2000) and Véronis (2001)) studies on the interannotator agreement report figures between 70% and 80% when WordNet is adopted as a sense inventory (Fellbaum *et al.* 1998; Edmonds and Kilgarriff, 2002; Snyder and Polmer 2004). An even worse figure of 67% is reported when sense choices are made by unknown contributors on the Web (Chklovski and Mihalcea 2003). Sense distinctions are rarely incontestable, and this is one of the reasons why we think it is important to refer to the lexicon structure as a justification for sense choices, as our method does.

A big dispute is even whether word senses exist at all (Hanks 2000; Kilgarriff 1997). Discretizing sense distinctions is difficult and can lead to different equally adequate choices. In contrast to the widespread enumerative approach, a generative approach has been proposed (Pustejovsky 1995) which consists of generating word senses based on the combination of a set of features called *qualia roles*. The latter approach overcomes the granularity problem by adopting a continuous representation of

senses, but presents other issues: it is unclear how to compare different system outputs in an objective manner when applied to a gold standard data set or how to consistently provide an unequivocal semantic annotation of text.

If we believe in word senses, the use of coarse-grained sense distinctions (or, alternatively, different levels of granularity) seems to be a promising approach to overcome the problem of low interannotator agreement. Ng, Lım and Foo (1999) show that, when a coarse-grained sense inventory is adopted, a consistent increase of the interannotator agreement is observed. Numerous manual and automated approaches to sense clustering have been proposed, ranging from syntactic and semantic criteria for grouping senses (Palmer 2000; Palmer, Dang, and Fellbaum 2007) to the use of heuristics (Peters, Peters, and Vossen 1998), dictionary mappings (Dolan 1994; Navigli 2006c), and confusion matrices (Chklovski and Mihalcea 2003; Agirre and de Lacalle 2003).

Recently, Hovy, Marcus, Palmer, Ramshaw, and Weischedel (2006) presented the OntoNotes project. The proposed approach aims at creating sense groupings: senses are iteratively partitioned until an interannotator agreement of 90% is reached in a sense annotation task. As a result the method does not need a further evaluation of its outcome, as it achieves its main objective, i.e. guaranteeing a high interannotator agreement. However, this approach – although faster than a classic manual sense creation procedure – still requires a great amount of work to produce a coarse set of senses on a large scale. Finally, an approach for clustering word senses based on lexico-syntactic features has been proposed which achieves state-of-the-art performance in WSD (Kohomban and Lee 2007).

Unfortunately, none of these approaches led to the release of an acknowledged *de facto* standard. As a result, until the WordNet sense inventory is not fully revised or a new sense inventory is widely adopted, enumerative approaches to coarse-grained WSD will remain unfeasible on a large scale. Until then, the most reasonable solution is to support annotators and, at a later stage, adjudicators in the difficult task of making the most appropriate choice from a discrete list of senses.

An important step in the direction of supporting sense annotations is the OMWE, a project for the acquisition of annotations with the aid of a web interface that presents the sense inventory of a word to be disambiguated in context (Chklovski and Mihalcea 2002). The main limitation of such an interface is that it is very difficult to provide an annotator, especially an unskilled one, with enough information to choose between subtle sense distinctions. In contrast, the method presented in this paper, even though applied to adjudication rather than sense annotation, has the advantage of supporting sense choices which are or seem coherent with respect to the adopted lexicon. A further difference is given by the fact that the OMWE expects a large number of contributions for the same item to be annotated reliably, whereas our approach can be applied to items annotated by any number of humans or systems. In our experiments we focused on two- or three-way disagreements; however, the very same approach can be employed when more annotators are involved at virtually no additional cost.

Other structural approaches, either based on semantic distances (e.g., Jiang and Conrath, 1997; Leacock, Chodorow, and Miller 1998) or graphs (e.g., Mihalcea

et al., 2004), could be employed to support the adjudication task. Unfortunately, knowledge-based methods are likely to perform poorly if they do not rely on rich lexical knowledge bases. This assertion is also supported by the findings discussed in the last part of Section 5.1.1. Also, most of these approaches cannot produce justifications for their choices in terms of semantic graphs as SSI does.

We stress that the adjudication task presented in this paper is only apparently easy: while it is true that the task works on a restricted number of senses for disagreed words, these are usually the hardest cases (i.e. those which make WSD an "AI-hard" task), simply due to the fact that humans or state-of-the-art systems disagreed on them.

As remarked in Section 2.2, ensemble approaches, i.e. combinations of WSD methods, are mostly related to the adjudication task. However, they differ in many aspects. Firstly, they do not exploit the agreed senses as an aid for solving disagreements. Secondly, they are applied to the outcome of automatic systems (often requiring a confidence degree to be output for each sense choice), in contrast to our approach which was applied to disagreements resulting from both manual and automatic annotation efforts. Thirdly, some ensemble approaches are supervised and require a training phase either for the disambiguation algorithms (first-order classifiers) or for the combination scheme (second-order classifier). This third aspect makes ensembles involving supervision unsuitable for the adjudication of disagreements resulting from the annotation of both unrestricted and domain-oriented texts.

Florian, Cucerzan, Schafer, and Yarowsky (2002) propose several kinds of combination approaches, based on voting, probability mixture, etc. Experiments are performed on the Senseval-2 lexical sample task: this kind of task is different from the one proposed in the present paper (an all-words task indeed), as it targets a single word per sentence (and neither requires nor exploits the annotation of the other words with senses). Klein, Toutanova, Ilhan, Kamvar, and Manning (2002) present ensemble approaches based on majority voting, weighted voting, and maximum entropy. The latter performs best (obtaining an improvement of 1% over the best single classifier), but requires a training phase on the second-order maximum entropy classifier. Experiments are again performed on a lexical sample basis, which makes a comparison with our approach not feasible.

Among the unsupervised approaches, purely relying on first- and second-order unsupervised classifiers, Stevenson and Wilks (2001) present a method for combining different disambiguation techniques, such as selectional preferences, subject codes, simulated annealing, etc. However, their evaluation is performed on a portion of the SemCor corpus, where each original sense tag has been mapped to an entry of the LDOCE dictionary. As a result, the authors obtain a performance around 90% and beyond 94% for fine- and coarse-grained disambiguation, respectively, which is difficult to compare to other systems, including ours.

Finally, we cite a study by Brody, Navigli and Lapata (2006), which concerns ensembles of unsupervised disambiguation algorithms. One of the combination approaches involves the use of SSI, the algorithm adopted in the present paper, as an arbiter of disagreed senses. Notice however that in that work the authors focus on

the adjudication of disagreed nouns. As a result, verbs, adjectives, and adverbs are left ambiguous and do not provide a strong contribution to the structural disambiguation of disagreed nominal senses. The other ensembles presented in the paper (majority voting, probability mixture, rank-based) would not benefit from agreed senses of other parts of speech, and this is the reason for the lower performance of SSI in that specific setting. In that work, SSI on the adjudication of disagreed nouns in SemCor obtains 56.3% versus 58.1% accuracy obtained by the best ensemble method, namely the rank-based approach (−1.8%). In contrast, in this paper we exploit the full power of sense-tagging agreements, thus obtaining results which overcome the state of the art in all-words WSD by around 3.3% (cf. Section 5.2).

In previous works (Navigli 2006a, b), we presented preliminary experiments on the manual and automatic adjudication of sense annotations. However, those experiments were inconclusive: on the manual side, we simulated disagreements on a much smaller scale without even distinguishing among classes of granularity as we did in this paper; on the automatic side, we performed experiments only on a part of the Senseval-3 test set, whereas in this paper we applied SSI to the full set of disagreements. All the experiments in the present paper are novel and provide interesting insights into the contribution of structural information to the resolution of disagreements.

7 Conclusions

In this paper we discussed the use of a specific kind of lexical chains, namely, structural semantic interconnections, to automatize the task of adjudicating manual and automatic disagreed sense assignments. The use of semantic interconnection patterns to support adjudication allows it to smooth possible divergences between the annotators and to corroborate choices consistent with the adopted lexicon. This novel aspect has been thoroughly assessed in simulated and real experiments, all based on standard data sets. The approach proves to be effective in the adjudication of human disagreements and exceeds by more than three points in accuracy to that of state-of-the-art all-words disambiguation systems with no additional training. The method is independent of the lexicon (i.e. WordNet), in that valid patterns can be derived from any sufficiently rich ontological resource, and can be applied to texts of any nature, in contrast to most combination methods.

The experiments illustrated in Section 5 show that semantic interconnections are a good means for providing coherent suggestions with a satisfactory balance between precision and recall (around 80% and 60%, respectively, on human disagreements). We remark the structural nature of the suggestions provided by SSI: a human judge can indeed visually analyze the correctness of an adjudication in terms of its semantic interconnections with respect to the other word senses chosen in context. The method has been implemented as a visual tool available online, called *Valido*[8] (Navigli 2006a). The tool takes as input a corpus of documents whose sentences are tagged by one or more annotators with word senses from the WordNet inventory. The user can

[8] Valido is available online at: `http://lcl.uniroma1.it/valido`.

browse the sentences, and adjudicate a choice over the others in case of disagreement among the annotators. In a future work, we plan to perform experiments on the usefulness of the visual suggestions provided by the tool in manually validating disagreed sense annotations, compared to a typical adjudication task performed in a nonvisual manner. We also aim at showing that semantic interconnections could be used during the annotation phase by taggers looking for suggestions based on the lexicon structure, with the objective of improving the coherence and awareness in the decisions to be taken. Finally, we would like to experiment on the use of WordNet domains (Magnini and Cavaglià 2000) to determine whether there is room for further improvement of our method in the automated resolution of disagreements.

Acknowledgments

This work is partially funded by the Interop NoE (508011), 6th EU FP. We wish to thank Francesco Maria Tucci, Giuliana and Maria Luisa Pesaresi, and Orin Hargraves for their invaluable support. We also thank the four anonymous reviewers who provided helpful suggestions to improve the paper.

References

Agirre, Eneko and de Lacalle, Oier López. 2003. Clustering wordnet word senses. In *Proceedings of Conference on Recent Advances on Natural Language (RANLP)*, Borovets, Bulgary, pp. 121–30.

Agirre, Eneko, Martínez, David, de Lacalle, Oier López, and Soroa, Aitor. 2006. Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, pp. 585–93.

Barzilay, Regina and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, pp. 10–17.

Bentivogli, Luisa, Pamela Forner, and Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the multisemcor corpus. In *Proceedings of the* 20*th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 364–70.

Berners-Lee, Tim. 1999. *Weaving the Web*. Harper, San Francisco, CA, USA.

Brody, Samuel, Roberto Navigli, and Mirella Lapata. 2006. Ensemble methods for unsupervised WSD. In *Proceedings of the* 44*th Annual Meeting of the Association for Computational Linguistics joint with the* 21*st International Conference on Computational Linguistics (COLING-ACL 2006)*, Sydney, Australia, pp. 97–104.

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

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

Cohen, Jacob A. 1960. A coefficient of agreement of nominal scales. *Educational and Psychological Measurement* **20**(1): 37–46.

Cuadros, Montse and German, Rigau. 2006. Quality assessment of large scale knowledge resources. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, pp. 534–41.

Decadt, Bart, Véronique Hoste, Walter Daelemans, and van den Bosch Antal. 2004. Gambl, genetic algorithm optimization of memory-based wsd. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*. Barcelona, Spain, pp. 108–12.

Dolan, William B. 1994. Word sense ambiguation: clustering related senses. In *Proceedings of 15th Conference on Computational Linguistics (COLING)*, Kyoto, Japan, pp. 712–16.

Edmonds, Philip and Adam, Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Journal of Natural Language Engineering* **8**(4): 279–91.

Fellbaum, Christiane (ed.) 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Fellbaum, Christiane, Joachim, Grabowski, and Shari, Landes. 1998. Performance and confidence in a semantic annotation task. In Christiane Fellbaum (ed.) *WordNet: an Electronic Lexical Database*, pp. 217–37, MIT Press, Cambridge, MA, USA.

Florian, Radu, Silviu Cucerzan, Charles Schafer, and David Yarowsky. 2002. Combining classifiers for word sense disambiguation. *Journal of Natural Language Engineering* **8**(4): 1–14.

Galley, Michel and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, Acapulco, Mexico, pp. 1486–8.

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities* **34**(1–2): 205–15.

Harabagiu, Sanda, George Miller, and Dan Moldovan. 1999. Wordnet 2 – a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX-99*, University of Maryland, USA, pp. 1–8.

Hirst, Graeme and St-Onge, David. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum (ed.) *WordNet: An electronic lexical database*, pp. 305–32, MIT Press.

Hovy, Eduard H., Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, New York, USA.

Jiang, Jay J. and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan, pp. 19–33.

Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* **31**(2): 91–113.

Klein, Dan, Kristina Toutanova, H. Tolga Ilhan, Sepandar D. Kamvar, and Christopher D. Manning. 2002. Combining heterogeneous classifiers for word-sense disambiguation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, Morristown, NJ, pp. 74–80.

Kohomban, Upali Sathyajith and Lee Wee Sun. 2007. Optimizing classifier performance in word sense disambiguation by redefining sense classes. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, pp. 1635–40.

Lea, Diana (ed.) 2002. *Oxford Collocations*. Oxford University Press, USA.

Leacock, Claudia, Martin Chodorow, and George Miller. 1998. Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics* **24**(1): 147–65.

Litkowski, Ken. 2004. Senseval-3 task: word-sense disambiguation of wordnet glosses. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain, pp. 13–16.

Longman (ed.) 2003. *Longman Language Activator*. Pearson Education, Harlaw, Essex, UK.

Magnini, Bernardo and Gabriela Cavaglià. 2000. Integrating subject field codes into wordnet. In *Proceedings of the 2nd Conference on Language Resources and Evaluation (LREC)*, Athens, Greece, pp. 1413–18.

Mihalcea, Rada and Ehsanul Faruque. 2004. Senselearner: minimally supervised word sense disambiguation for all words in open text. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain, pp. 155–8.

Mihalcea, Rada, Paul Tarau, and Elizabeth Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the* 20*th COLING 2004*, Geneva, Switzerland, pp. 1126–32.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, Princeton, NJ, USA, pp. 303–8.

Miller, Irwin and Marylees Miller (eds.) 2003. *John E. Freund's Mathematical Statistics with Applications,* 7*th Edition*. Prentice Hall, NJ, USA.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* **17**(1): 21–43.

Navigli, Roberto. 2005. Semi-automatic extension of large-scale linguistic knowledge bases. In *Proceedings of the* 18*th FLAIRS*, Clearwater Beach, USA, pp. 548–53.

Navigli, Roberto. 2006a. Consistent validation of manual and automatic sense annotations with the aid of semantic graphs. *Computational Linguistics* **32**(2): 273–81.

Navigli, Roberto. 2006b. Experiments on the validation of sense annotations assisted by lexical chains. In *Proceedings of the European Chapter of the Annual Meeting of the Association for Computational Linguistics (EACL)*, Trento, Italy, pp. 129–36.

Navigli, Roberto. 2006c. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the* 44*th Annual Meeting of the Association for Computational Linguistics joint with the* 21*st International Conference on Computational Linguistics (COLING-ACL 2006)*, Sydney, Australia, pp. 105–12.

Navigli, Roberto 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): 1075–88.

Navigli, Roberto, Kenneth C. Litkowski, and Orin Hargraves. 2007. Semeval-2007 task 07: coarse-grained english all-words task. In *Proceedings of the* 4*th International Workshop on Semantic Evaluations (SemEval-2007)*, Czech Republic, pp. 30–5, Prague, Association for Computational Linguistics.

Ng, Hwee T., Chung Y. Lim, and Shou K. Foo. 1999. A case study on the inter-annotator agreement for word sense disambiguation. In *Proceedings of ACL Workshop: Standardizing Lexical Resources*, College Park, MD, pp. 9–13.

Palmer, Martha. 2000. Consistent criteria for sense distinctions. *Computers and the Humanities* **34**(1–2): 217–22.

Palmer, Martha, Hoa Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering* **13**(2): 137–63.

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

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, pp. 21–5.

Pustejovsky, James. 1995. *The Generative Lexicon*. Cambridge, MA, MIT Press.

Rigau, German, Jordi Atserias, and Eneko Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of* 35*th Annual Meeting of the Association for Computational Linguistics joint with* 8*th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL'97)*, Madrid, Spain, pp. 48–55.

Snyder, Benjamin and Martha Palmer. 2004. The english all-words task. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain, pp. 41–43.

Soanes, Catherine and Angus Stevenson (ed.) 2003. *Oxford Dictionary of English*. Oxford University Press.

Stevenson, Mark and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics* **27**(3): 321–49.

Véronis, Jean. 2001. Sense tagging: does it make sense? In *Corpus Linguistics 2001 Conference*, Lancaster, UK.

Véronis, Jean. 2004. Hyperlex: lexical cartography for information retrieval. *Computer, Speech and Language* **18**(3): 223–52.

Yuret, Deniz. 2004. Some experiments with a naive bayes wsd system. In *Proceedings of ACL 2004 SENSEVAL-3 Workshop*, Barcelona, Spain, pp. 265–68.